Journal of Biomedical Informatics 51 (2014) 100–106

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Improving search over Electronic Health Records using UMLS-based query expansion through random walks

David Martinez [a,*], Arantxa Otegi [b], Aitor Soroa [b], Eneko Agirre [b]

[a] CIS Department, University of Melbourne, Melbourne 3010, Australia
[b] IXA NLP Group, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, Donostia, Basque Country, Spain

ARTICLE INFO

Article history:
Received 20 August 2013
Accepted 12 April 2014
Available online 21 April 2014

Keywords:
Information storage and retrieval
Algorithms
Data mining
Semantics
Natural language processing

Objective: Most of the information in Electronic Health Records (EHRs) is represented in free textual form. Practitioners searching EHRs need to phrase their queries carefully, as the record might use synonyms or other related words. In this paper we show that an automatic query expansion method based on the Unified Medicine Language System (UMLS) Metathesaurus improves the results of a robust baseline when searching EHRs.

Materials and methods: The method uses a graph representation of the lexical units, concepts and relations in the UMLS Metathesaurus. It is based on random walks over the graph, which start on the query terms. Random walks are a well-studied discipline in both Web and Knowledge Base datasets.

Results: Our experiments over the TREC Medical Record track show improvements in both the 2011 and 2012 datasets over a strong baseline.

Discussion: Our analysis shows that the success of our method is due to the automatic expansion of the query with extra terms, even when they are not directly related in the UMLS Metathesaurus. The terms added in the expansion go beyond simple synonyms, and also add other kinds of topically related terms.

Conclusions: Expansion of queries using related terms in the UMLS Metathesaurus beyond synonymy is an effective way to overcome the gap between query and document vocabularies when searching for patient cohorts.

1. Introduction

The automatic processing of Electronic Health Records (EHRs) offers exciting possibilities for applications such as phenotyping and biosurveillance. One of the main obstacles to exploit EHRs arises from the fact that large portions of information are encoded as free text. Thus, techniques from Natural Language Processing (NLP) and Information Retrieval (IR) are necessary, and this has stimulated a wealth of research in this area. Previous research has studied different problems in this domain, such as acronym and abbreviation resolution in clinical discourse [20], finding cancer staging information in text [14], or mapping references in text into a medical ontology [1]. Recently, the project Khresmoi[1] has created a large consortium of European Research Centers joining forces to improve medical analysis information and retrieval.

However, the datasets used in most of these projects have not been shared with the research community, often due to ethical clearance issues, and this has made the comparison of different methodologies difficult. Some recent initiatives have tried to address the lack of shared EHR testbeds, creating NLP and IR challenges where different systems participate. The most active has been the i2b2 (Informatics for Integrating Biology and the Bedside) initiative, funded by the National Institute of Health (NIH).[2] This center has organised several challenges since 2006, involving the following NLP tasks: de-identification of reports; classification of smoking status; recognition of obesity and co-morbidities; extraction of medication information; extraction of concepts, assertions and relations; co-reference identification; and temporal relations in EHR.

The study of IR over EHR information sources has been even less explored than other NLP-related tasks, due to the need of larger collections and the complexity involved in building evaluation datasets, which include the selection of appropriate queries, the need of relevance judgements, etc. The main contribution to this line of research came with the 2011 and 2012 Medical Records

* Corresponding author.
E-mail addresses: davidm@csse.unimelb.edu.au (D. Martinez), arantza.otegi@ehu.es (A. Otegi), a.soroa@ehu.es (A. Soroa), e.agirre@ehu.es (E. Agirre).

[1] http://www.khresmoi.eu/overview/.

[2] https://www.i2b2.org/index.html.

http://dx.doi.org/10.1016/j.jbi.2014.04.013
1532-0464/© 2014 Elsevier Inc. All rights reserved.

tracks at the Text Retrieval Conference (TREC).[3] For the first time, search for cohorts of patients for relevant medical queries was attempted at a large scale (more than 100,000 medical records). The queries were built by targeting a list of research areas that the U.S. Institute of Medicine considered priorities for comparative effectiveness research.[4] The relevance assessment was done by groups of clinicians after pooling documents for each query. The queries included different pathologies and treatments, as well as demographic constraints. These challenges allowed the comparison of different systems over a shared dataset. They attracted the interest of 29 research groups in 2011 and 24 groups in 2012.

In this paper we present in detail a random walk approach for IR over EHR, which performs automatic query expansion based on the concepts and relations in the knowledge bases included in the Unified Medical Language System (UMLS). The query expansion method relies on an algorithm based on random walks, known as Personalised PageRank, which is run over a graph representation of the UMLS. The intuition behind our approach is the following: if we initialise the probability distribution of the UMLS graph with the terms identified in the query, the random walk will help identify relevant terms, which can be used to expand the query for improved retrieval. Our approach is flexible with regard to the type of relationship, and a variety of related terms are found with this method. For instance, for the query "Patients with Primary Open Angle Glaucoma (POAG)", terms such as "eye" and "ocular" are selected by the system for expansion, leading to improved retrieval performance.

This article is partially based on the results obtained in our participation in the TREC Medical Track in 2012, as reported in the TREC working notes [15]. Our participation combined several well-known query expansion techniques with a method based on random walks. In this article we provide a better framing of our proposed algorithm, as well as additional analysis, which allows to separate the contribution of our expansion algorithm from the contributions of the other techniques.

The PageRank graph-based random walk algorithm was first introduced by Page et al. [17] as a way to better represent the topology of the WWW in order to improve search. Since then, it has been used for a variety of problems, including the prediction of gene markers for cancer [24]. Personalised PageRank [9] was developed in order to represent the importance of a particular query when initialising the probability distribution, and it has been successfully used in NLP tasks such as Word Sense Disambiguation (WSD) [5,4,6,21] and word similarity [19,2,3]. It has been applied both to a general purpose lexical knowledge-bases such as WordNet [2,3,5,4] and also to the UMLS [6,21]. In addition, recent results show that it is useful to improve ad hoc IR with WordNet [16]. In this work, we introduce a method to apply Personalised PageRank over the UMLS to the EHR retrieval tasks of TREC 2011 and 2012. We first present a baseline system, which achieved strong results in the competition in both years, and then show that our query expansion technique yields improvements over the baseline on both datasets.

2. Material and methods

In this section we first present the document and query collections used in the Medical Records track organised by TREC in 2011 and 2012. This shared task was an ad hoc search challenge that modelled the clinical task of finding patient cohorts for comparative effectiveness research [23]. Next, we describe the steps of our method, including the processing of the documents and queries, as used in the baseline approach. We then present our query expansion technique and, finally, we detail the indexing and searching steps.

2.1. Document and query collections

The document set was almost the same in both editions, with minor changes in the collection for the TREC-2012 challenge, which contains 844 fewer reports than TREC-2011. The set consists of de-identified clinical reports made available for the competition through the University of Pittsburgh. It contains one month of reports from multiple hospitals, and includes nine types of reports, such as Radiology Reports, Consultation Reports, and Surgical Pathology Reports. In addition to free text, the reports also include several ICD codes.[5] Each report is linked to a "visit", which represents a patient's single stay at a hospital. The set contains around 100,000 reports, which are grouped in around 17,000 visits. The unit of retrieval in the track was the visit, i.e. the union of the content of all the reports associated with that visit.

The test set for TREC-2011 contains 34 topics (or queries), and TREC-2012 consists of 50 topics. The topics were obtained by relying on two sources: (i) a list of research areas that the U.S. Institute of Medicine (IOM) had deemed priorities for clinical comparative effectiveness research,[6] and (ii) the OHSUMED literature retrieval test collection. Some example topics are given below:

- Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis.
- Patients with hearing loss.
- Female patients with breast cancer with mastectomies during admission.
- Adult patients who received colonoscopies during admission which revealed adenocarcinoma.

The relevance assessments were collected manually, with several groups of domain experts going over the sets of documents pooled from the runs of participating systems. For the year 2011 there were two submission deadlines, and only the systems that met the first deadline contributed to the pool (47). In 2012 there was a single deadline, and all 88 runs were used for pooling.

2.2. Processing the document collection

In order to explore the benefits of the random walk algorithm over a strong baseline, we rely on the pipeline developed by Karimi et al. [11] to process the document collection. The pipeline distinguishes several fields based on patterns and medical codes as found in the document collection, and allows to build several indices. This pipeline participated in TREC-2011 with competitive results, and was used to build our baseline system. The process consists on the four steps listed below, which are combined in different ways to build document indices, as described in Section 2.5. The document processing steps are the following:

- Expansion of mentions of ICD9 codes into text descriptions.
- Identification of headings.
- Identification of negations.
- Stemming.

The first step expands the mentions of ICD9 codes of admission and discharge diagnoses in the metadata with their text

[3] http://trec.nist.gov/pubs/trec20/t20.proceedings.html and http://trec.nist.gov/pubs/trec21/t21.proceedings-tracks.html#medical.

[4] http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx.

[5] International statistical classification of diseases and related health problems: http://en.wikipedia.org/wiki/List_of_ICD-9_codes.

[6] http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.

descriptions. Both the original code and expanded forms are included for indexing.

In the second step, heading identification is performed. The documents in the collection contain different sections, with their corresponding headings. The system applies hand-crafted pattern-matching rules to identify the main headings, in order to build different indices and allow for field-based search. In addition, hand-made rules are also used to identify and normalise some demographic information, as, for instance, gender, age, and other specific conditions (such as weight) mentioned in the text, so it is added to separate fields. The extracted fields are the following:

- ADMITDIAG: Diagnostic during admission.
- AGE: Patients age by decades (e.g. age30 means people in their thirties).
- ALLERGIES: Allergies listed in the report.
- CHIEFCOMP: Chief complaint, this may be equal to diagnostics during admission.
- DISCHDIAG: Discharge diagnostics.
- GENDER: Patient's gender extracted from text.
- HISTORY: History of the patient's medical condition or past medical illness.
- MEDICATIONS: Medications.
- PRESHIS: Present illness medical history.
- PASTHIS: Past medical history.
- REPORT: All the textual information, including history (past and present).

The third step runs NegEx[7] over the entire collection in order to detect negated phrases. It relies on the NegEx parser built into MetaMap-2010 [7], which specifies which of the identified phrases are negated. We use this information to build an index that converts negated terms into an encoded representation whenever the negated form is the most frequent in the document. For instance, when a document in the collection contains a sentence such as "There is no chronic back pain", NegEx detects that negation is implied for the phrase "chronic back pain". After parsing the whole document, if "chronic back pain" appears in negated form more often than in positive form, all instances of "chronic back pain" in the document are replaced with the word "nochronicbackpain", that is, the negated phrase is transformed into a single word, with no spaces, and a "no" prefix.

Our aim with the negated index is to avoid matching cases where the term is framed as negative in the document more often than as positive. Due to the lack of negated queries in the collections, the result of transforming words is the same as removing them. The method to detect negation allows us to build indices with and without this module.

Finally, the fourth step is to stem the words in the documents using the Porter stemmer. We experimented with both stemmed and original words.

2.3. Processing queries: identifying fields

Based on Karimi et al. [11], we apply a set of manually constructed patterns to map query terms into the available fields. These patterns were based on the sample clinical questions provided by the National Library of Medicine (NLM) as provided by [10]; and covered seven broad categories of age, weight (using body mass index), diagnostics, treatments, medications, history, allergies and abbreviations. For example, if a query contained "elderly patients", we expanded "elderly" with an equivalent age field that covered people in their 60s to 90+. Table 1 shows all the selected transformation rules. For example the query:

*Elderly patients with ventilator-associated pneumonia*is automatically translated into:

PRESTHIS: (ventilator associated pneumonia) OR
DISCHDIAG: (ventilator associated pneumonia) OR
AGE: (age60 age70 age80 age90) OR
REPORT: (elderly with ventilator associated pneumonia).

2.4. Query expansion using Personalised PageRank

In addition to the processing mentioned in the previous section, which is the core of the baseline system, we propose to use an automatic method for query expansion. For this, we use a graph algorithm based on random walks over the graph representation of a knowledge-base of concepts and relations, which yields terms related to the input query. The UMLS Metathesaurus is used as the knowledge-base, and we thus represent the UMLS as a graph.

The UMLS Metathesaurus contains a wide range of information about the relations between terms in the form of database tables. The MRREL table lists relations between concepts, such as "parent", "can be qualified by" and "related and possibly synonymous" among others. In order to obtain the graph structure of the UMLS, we simply treat the concepts in the UMLS as vertices, and the relations listed in the MRREL table as directed edges. No weights are used for the relations that are extracted from the MRREL table. The graph construction method is the same as in Agirre et al. [6], and uses publicly available scripts[8] on version 2012AA of the UMLS.

We did not explicitly represent and use the rich information on types and terminology sources of relations and concepts. In related work on WSD using UMLS [6], the authors did an analysis of relations and sources, and the results showed that using all the relations and terminology sources was beneficial, with no explicit representation of them.

Given a query and the graph-based representation of the UMLS, we obtain a ranked list of related concepts as follows:

1. We first identify the fields in the query (cf. Section 2.3), and then run MetaMap to identify the UMLS terms in the query. MetaMap allows to disambiguate those terms and return directly the relevant concept, using the built-in WSD module. We explored MetaMap both with and without WSD. Note that in the cases where the fields are recognised, we expand the value of each field separately.
2. We assign a uniform probability distribution to the concepts found in the query. The rest of the nodes are initialised to zero. If WSD is not turned on, we use all possible concepts meant by the term. For instance, for the term "cold", the three concepts in MetaMap will be assigned a uniform distribution of 1/3: C0009443 "Common Cold", C0009264 "Cold Temperature" and C0234192 "Cold Sensation". When WSD is turned on, only the concept returned by MetaMap is used. In the case of "cold", and depending of the context in the query, one of the concepts just mentioned would be used.
3. We compute Personalised PageRank [9] over the graph (see below). The computation is initialised using the concepts in the query, as mentioned in the previous step. The result is a probability distribution over UMLS concepts. The higher the probability for a concept, the more related it is to the given query.

The intuition behind Personalised PageRank is that of an agent performing infinite walks in the graph at random. The agent starts

[7] http://code.google.com/p/negex.

[8] http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/biomedical_wsd/.

Table 1
Rules (patterns in the queries and their translations) used in the query transformation step. Words that are all in capital letters are field names (cf. Section 2.2).

What	Pattern	Translation
Gender	Women/female	GENDER:genderfemale
	Men/male	GENDER:gendermale
Age	Young adult	AGE: (age20 age30 age40)
	Younger/young	AGE: (agebirth12 ageteen age20 age30 age40)
	Adult	AGE: (age20 age30 age40 age50 age60 age70 age80 age90)
Weight	(BMI|Body Mass Index)	
	(Bigger than|more than|of|approximately of)	
	>= 36	WEIGHT: (obesity obese overweight “morbidly obese” “morbid obese” “morbid obesity” “markedly obese”)
	>=30 and <=35	WEIGHT: (obesity obese overweight “moderately obese” “moderate obesity”)
	>=25 and <=30	WEIGHT: (obesity obese overweight “slightly obese” “mildly obese”)
	>=18.5 and <=25	WEIGHT: (“normal weight”)
	(BMI|Body Mass Index)	
	(less than|of|approximately of)	
	>16 and <=18.5	WEIGHT: (underweight)
	<=16	WEIGHT: (underweight “severely underweight”)
Treatments	Taking X (who|with|without|treated)	MEDICATIONS:X
	Who are on X	MEDICATIONS:X
	Patients on X for Y	MEDICATIONS:X
Admission	Admitted (for|with) X who	CHIEFCOMP:X OR ADMITDIAG:X
Diagnostics	Treated for X (who|during|while)	PRESTHIS:X OR DISCHDIAG:X
	(Patients with|men with|women with) X	PRESTHIS:X OR DISCHDIAG:X
	Who were discharged X	DISCHDIAG:X
History	With a* history of X (who|now)	HISTORY:X
Allergy	With X allergy	ALLERGY:X
	Without allergy	ALLERGY: (noallergies)
‘ER’ abbreviation	Seen in the er|presented to the er	REPORT: (“emergency room” OR ER)

in any of the concepts mentioned in the query, and follows at random one of the relations to another concept, then to another concept, ad infinitum. With certain probability, the agent would restart in any of the concepts mentioned in the query, and continue its walk. If the agent walked infinite time, the number of visits to each concept in the graph would give an indication of how related that concept is to the query terms. Traditional PageRank is based on the same intuition, but there is no reference to any query: instead of (re) starting the walk on specific concepts, it (re) starts in any concept in the knowledge-base, and it thus gives an indication of which concepts are more “central” in the graph, regardless of the query.

The implementation of Personalised PageRank is based on the traditional PageRank equation, which we formalise as follows. Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree[9] of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if an edge from i to j exists, and zero otherwise. The transition matrix contains all relations in MRREL. The calculation of the *PageRank vector* $\mathbf{P}$ over G is equivalent to resolving Eq. (1).

$$\mathbf{P} = cM\mathbf{P} + (1-c)\mathbf{v} \tag{1}$$

The first term of the sum in the equation models the case where the random walker follows one of the relations. In the equation, c is the so-called *damping factor*, a scalar value between 0 and 1, which models the probability of the random walker following one relation, instead of restarting. The second term represents restarting, loosely speaking, the probability of the random walker jumping to any node, i.e. without following any relation in the graph. $\mathbf{v}$ models the concepts where the random walker might restart, and is a $N \times 1$ vector of probabilities. The damping factor, usually set in the $[0.85 \ldots 0.95]$ range [9], models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that the PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalised vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of Personalised PageRank as used here, $\mathbf{v}$ is initialised with uniform probabilities for the concepts in the query, and 0 for the rest of concepts. In other words, Personalised PageRank is computed by modifying the random walk distribution vector in the traditional PageRank equation. As said above, all probability mass is concentrated on the concepts identified in the query.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (1) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation.[10] Following usual practice, we used a damping value of 0.95, and 30 iterations.

As an illustration of the Personalised PageRank scores, Fig. 1 shows a subgraph of the UMLS around concept “Primary open glaucoma” (C0339573). If this concept was found in a query, we would (re) start the random walk distribution on that concept, and the concepts around it would receive high scores. As the Figure shows, the top ranking concepts would include “Glaucoma syndrome” (C0017601), “Eye, optic” (C0015392) and “GLC1E” (C1842026).

In order to select the expansion terms for a given query, we sort all concepts in the UMLS according to their Personalised PageRank value, and pick the top concepts according to a threshold. The selected concepts are then lexicalised, using the terms specified in the UMLS Metathesaurus. We explored two approaches to select the top concepts: (i) select the top k concepts, or (ii) select all the concepts with weights above a given threshold t. Our preliminary experiments over the TREC-2011 dataset suggested that the former approach was able to provide better performances for different settings, and we thus decided to use the top k concepts for our

[9] The number of edges starting in edge i.

[10] http://ixa2.si.ehu.es/ukb/.

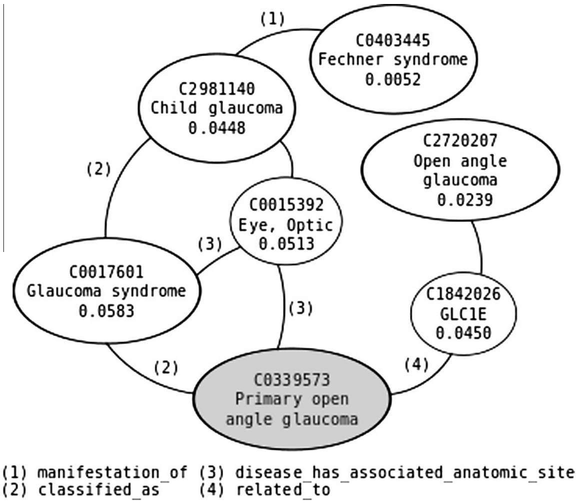

Fig. 1. Example showing a partial subgraph of UMLS centred around the concept "Primary Open Angle Glaucoma" (C0339573). Relation types are also described when applicable as UMLS does not provide type labels for all the relations. The numbers in the nodes correspond to their rank after computing Personalised PageRank, where the random walk distribution is concentrated on the concept C0339573 (in grey).

experiments, where k was selected according to development experiments (see Section 3).

2.5. Indexing and searching

In this section we describe the method for indexing all documents and performing the search. We explored several alternatives of the baseline and our query expansion technique. As a first parameter, we distinguish between two types of indexing in our runs: visit-based and report-based. In the former approach, all related reports for a visit were concatenated (removing duplicate diagnostics codes) to create a single "multi-document" item for indexing. In the latter, each report is indexed as a separate item. We refer to the former approach as VISIT, and as REPORT to the latter.

As explained before, we extracted fields from the query and generated different indices depending on the use of separate fields or not (FIELDS/COMBINED). The COMBINED index simply represents the normalised terms as a flat string. The use of fields (FIELDS) is implemented as a Boolean search over the fields, where the operator 'OR' is used to join the results on each field index, and a ranking of documents is obtained from all the documents retrieved.

Another alternative we explored refers to the application (or not) of stemming when indexing (STEM/NOSTEM). When using stemming, we performed stop-word removal both in query processing and indexing. The standard list of stop-words was augmented with the word *patient*, and we also removed all single characters and words *and*, *or*, *not*, and *no* from the list.

As mentioned before, there were minor changes between the document collections of TREC-2011 and TREC-2012 (a small amount of reports were not included in 2012), and we generated separate indices for each query set. Regarding negation, we preprocessed the document collection with NegEx, in order to handle negated terms, and built separate indices (cf. Section 2.2). However, few terms were affected, and the changes were minimal when we tested the different indices over the 2011 dataset. We report the results for the NegEx-processed index for TREC-2011, and the full index for TREC-2012.[11]

The search engine used for indexing and searching in our runs was Apache Lucene (v3.2). We applied both the BM25 and the tf-idf ranking algorithms [18].

3. Results

The metric used in the experiments presented here is Bpref, the main evaluation metric of the 2011 TREC Medical Record track.[12] It was chosen because of its robustness for incomplete judgement sets, since it is computed on the basis of judged documents only [8]. It is inversely related to the fraction of judged non-relevant documents that are retrieved before judged relevant documents:

$$\text{Bpref} = \frac{1}{R}\sum_{r} 1 - \frac{|n\,\text{ranked higher than}\,r|}{R} \qquad (2)$$

where R is the number of documents judged relevant for a topic, r is a relevant retrieved document and n is a member of the first R retrieved documents judged non-relevant.

As explained before, we tested alternative configurations of the indexes. In summary, these are the different settings we explored:

- Use visit-based (VISIT) or report-based (REPORT) index.
- Application of stemming or not (STEM/NOSTEM).
- Use index with separate fields (FIELDS) or all fields together (COMBINED).
- Perform (or not) WSD prior to choosing the initial concepts when applying PageRank.
- Use either tf-idf or BM25 as the ranking algorithm.
- Process the query (following the steps to identify fields presented in Section 2.3) before or after applying PageRank (BEFORE/AFTER).
- Different thresholds (THR) for the number of top concepts to use for expansion, ranging from 3 to 20.

All alternatives apply to the baseline and our query expansion system, except for the last two, which only apply to our query expansion system.

We performed several experiments on each query set to select the best configuration. In order to avoid overfitting, the single best configuration of the TREC-2011 query set was used to test the TREC-2012 query set; and in the same way, we tested the TREC-2011 query set with the best configuration obtained on the TREC 2012 query set.

The best configurations for each development query set, including the baseline system and the PageRank expansion system, are listed in Table 2. We observe that the optimal setting for the baseline varies across the two datasets. The best configuration for the PageRank expansion system is quite consistent in both datasets: REPORT, COMBINED and TFIDF for indexing, not performing WSD, parsing the query AFTER PageRank, and a threshold of 3 or 4 concepts for expansion, although stemming seems to help in one dataset, and not in the other. The results on the development set showed that, in both cases, the use of PageRank expansion improved the results over the baseline.

The main experiments were performed in each query set using the best configuration according to the respective development set. Table 3 reports the results, showing that our expansion strategy based on PageRank improves over the baseline in both cases. The difference with the baseline is statistically significant according to the 2-tailed Student t-test ($p < 0.01$) for the TREC-2011 query set. The table also contains the best and mean scores among TREC participants. Our baseline systems are close to the mean of the TREC participants for both datasets, while the PageRank system

[11] We did not perform the NegEx processing step over the new 2012 dataset, due to the lack of impact over the 2011 dataset.

[12] The 2012 challenge relied also on inferred metrics.

Table 2
Best configurations for each system and development query set.

Development query set	System	Best configuration
TREC-2011	Baseline PageRank	VISIT + STEM + COMBINED + TFIDF REPORT + NOSTEM + COMBINED + TFIDF + NOWSD + AFTER + THR3
TREC-2012	Baseline PageRank	REPORT + STEM + FIELDS + BM25 REPORT + STEM + COMBINED + TFIDF + NOWSD + AFTER + THR4

Table 3
Results over the test query sets as Bpref.

Test collection	TREC systems		Our systems	
	Best	Mean	Baseline	PageRank
TREC-2011	0.5523	0.4283	0.4160	0.5469[a]
TREC-2012	0.4515	0.3288	0.3205	0.3542

[a] Indicates statistically significant improvement over the baseline.

Table 4
Queries with highest improvement for PageRank, together with the learnt expansion terms and the Bpref increase.

Query	TREC version	Expansion terms	Bpref increase
Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis	2011	MRSA elsewhere/NOS Personal history of poliomyelitis Personal history of other infectious and parasitic disease	0.931
Patients with Primary Open Angle Glaucoma (POAG)	2012	Eye, Eyeball, Globe, Ocular… Glaucoma syndrome Open cleft glaucoma GLC1E	0.742
Patients with adult respiratory distress syndrome	2012	Immunology Taxonomy Metabolism Historical aspects	0.722

is better in both, and close to the top system for the TREC-2011 collection. Our query expansion system would have ranked 3rd among the 109 automatic runs submitted to TREC-2011, and 20th among 82 automatic runs in TREC-2012,[13] which shows its competitiveness with the state of the art. The overall performance in TREC-2012 for our system is lower than in TREC 2011, but also for the rest of participants at the challenge, with the top automatic system reaching a Bpref of 0.4515 (compared to 0.5523 for the best system in TREC-2011).

In fact, the task coordinator mentions that the "2012 task was inherently harder" [22, p. 245], because the selection of topics in 2012 could have been biased against "obviously easy topics" [22, p. 241]. She also mentions that a specific participant [13] ran the same system on both years, finding that the system would have scored among the five top participants in 2011, but slightly above the median on the second, with a drop of 16 absolute points in Bpref between both years. This confirms that the 2012 topics were more difficult, and largely agrees with the results of our baseline.

Another aspect to consider are the differences between the best performing systems in the two challenges. In 2011 there was no training data available, and the top performing system [12] only relied on their own manual annotation to tune their set of parameters. They applied a document-analysing pipeline that built different indices depending on the different sections and types of the documents in the collection. This team did not participate in 2012, when the best system applied an ensemble model which benefited from combining several IR systems and expansion sources, trained over the full 2011 query set [26]. They performed cross-validation experiments over a large set of configurations, and their results suggest that their strategy performs well when using a training collection with a similar set of queries. Their earlier participation in 2011 [25] applied a single underlying model (mixture relevance model), and they estimated the parameters by concept overlapping of topics with external collections, instead of training data. Their best Bpref score in 2011 was 0.522 (below our PageRank system, cf. Table 3). Even after they developed a more sophisticated architecture for 2012 and achieved the best score, their absolute Bpref score was lower than in 2011.

The goal of our research is to show that query expansion beyond synonymy (based on the UMLS and PageRank) improves the results of a strong system, which we build using a single IR model. Surpassing the state of the art was not the goal of this investigation. We are aware that an ensemble of IR systems finely tuned on the 2011 data (as done by the top system in 2012) performs better, and we would like to explore incorporating our query expansion technique there in the future. In any case, note that our system has very few parameters and variations, making it, in principle, more robust to datasets which are not similar to the train data, as shown in related work [16,13]. We leave the exploration of these issues for the future.

4. Discussion

In order to analyse the reason for the improvement obtained with query expansion, we selected the queries where the difference in performance of PageRank with respect to the baseline was largest. Table 4 shows the UMLS concepts that were returned by our method, as used to expand some of those queries. We can see that the proposed expansion terms have different characteristics. Some terms are synonyms of a query term, e.g. "open cleft glaucoma" in the second query in Table 4 is synonymous to "open angle glaucoma". This was expected, as previous work has shown that synonyms are good candidates for query expansion. We also observe some high-level concepts that have a distant taxonomic relationship with the query terms, such as the expansion terms "metabolism" and "historical aspects" in the third query.[14] For the first row, the term contributing most to the performance gain is "Personal history of other infectious and parasitic disease", mainly because of "parasitic" which is related to topic of the question. For the third row, "Historical aspects" contributes to the improved performance by boosting the score of patients with mentions of "patient history" or "family history". The latter result was an artifact of the dataset, and not an intended effect.

Most interestingly, the examples also show that some relations beyond synonymy and taxonomy are helpful, e.g. the query term "glaucoma" in the second query is related to the expansion terms "eye" and "optic", according to a relationship of type "disease has associated anatomic site" in the UMLS (cf. Fig. 1). By looking at a sample of relevant documents, we observed that the expansion terms related to "eye" and "optic" are useful to boost the score

[13] Note that our official participation on TREC-2012 did not test PageRank-based expansion alone, as it combined several query expansion techniques.

[14] These terms occur as direct, unlabeled links in the UMLS graph.

of those documents because the glaucoma patients usually use medication to treat their eyes, and this is explicitly mentioned in most of the documents associated to them. All in all, the fact that they help to reach high performance in the cohort retrieval task is an indication that there is some correlation for those concepts over the positive patients in the data.

Our overall results over the two datasets suggest that our PageRank-based query expansion performs robustly in the task of retrieving patient cohorts. Another advantage of our method (and other explicit query expansion techniques) is that the added terms can be presented to the user in an interactive way for selection (such as in the PubMed interface for searching the biomedical literature); this would allow flexibility for building queries for finding patient cohorts.

5. Conclusions and further work

Most of the information in Electronic Health Record is represented in free textual form. Practitioners searching EHRs need to phrase their queries carefully, as the text in EHR might use synonyms or other related words instead of the query terms. We have presented an automatic query expansion method based on the UMLS, which improved the results of a strong baseline when searching for patient cohorts. The method uses a graph representation of the lexical units, concepts and relations in the UMLS Metathesaurus. It is based on random walks over the graph, where the random walks are initialised with the query terms.

Our experiments over the TREC Medical Record track show improvements in both the 2011 and 2012 datasets over a strong baseline. Note that we tuned the parameters of our method and the baseline on the other dataset. Our analysis shows that the success of our method is based on the automatic expansion of the query with different types of concepts. The terms added in the expansion go beyond simple synonyms, and they are helpful to identify patients that are considered relevant for the queries.

For the future, we would like to explore whether our query expansion technique would improve the results of a state-of-the-art system, both in the scenario where the test set is similar to the train set, and in the scenario where they differ [16]. Another avenue of research is the exploitation of the rich relation and knowledge sources of information in the UMLS.

Acknowledgments

This work was performed while David Martinez was employed at NICTA. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme. This research was partially funded by the Spanish Ministry of Economy (SKATER and READERS CHIST-ERA projects).

References

[1] Nguyen A, Lawley M, Hansen D, Colquist S. A simple pipeline application for identifying and negating snomed clinical terminology in free text. In: Health informatics conference, Canberra, Australia; 2009. p. 188–93.

[2] Agirre E, Alfonseca E, Hall K, Kravalova J, Paşca M, Soroa A. A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics. NAACL '09. Association for Computational Linguistics; 2009a. p. 19–27.

[3] Agirre E, Cuadros M, Rigau G, Soroa A. Exploring knowledge bases for similarity. In: Chair NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, editors. Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta; 2010a.

[4] Agirre E, de Lacalle OL, Soroa A. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In: Proceedings of IJCAI. Pasadena, USA; 2009b. p. 1501–6.

[5] Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation. In: Proceedings of the 12th conference of the European chapter of the association for computational linguistics. EACL '09. Association for Computational Linguistics; 2009. p. 33–41.

[6] Agirre E, Soroa A, Stevenson M. Graph-based word sense disambiguation of biomedical documents. Bioinformatics 2010;26(22):2889–96.

[7] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc JAMIA 2010;17(3):229–36. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995713&tool=pmcentrez&rendertype=abstract>.

[8] Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. Sheffield, UK; 2004. p. 25–32.

[9] Haveliwala TH. Topic-sensitive PageRank. In: Proceedings of WWW '02; 2002. pp. 517–26.

[10] Institute of Medicine. 100 initial priority topics for comparative effectiveness research; 2009. <http://www.iom.edu/~/media/Files/Report%20Files/2009/ComparativeEffectivenessResearchPriorities/CER%20report%20brief%2008-13-09.ashx>.

[11] Karimi S, Martinez D, Ghodke S, Zhang L, Suominen H, Cavedon L. Search for medical records: NICTA at TREC 2011 medical track. In: Proceedings of the text retrieval conference (TREC); 2012.

[12] King B, Wang L, Provalov I, Zhou J. Cengage learning at trec 2011 medical track. In: TREC; 2011.

[13] Leveling J, Goeuriot L, Kelly L, Jones GJ. DCU@TRECMed 2012: using ad-hoc baselines for domain-specific retrieval. In: Medical records track, text retrieval conference (TREC) 2012; 2012. pp. 1–9.

[14] Martinez D, Li Y. Information extraction from pathology reports in a hospital setting. In: ACM international conference on information and knowledge management (CIKM); 2011. p. 1877–82.

[15] Martinez D, Otegi A, Agirre E. NICTA and UBC at the TREC 2012 medical track. In: Proceedings of the text retrieval conference (TREC); 2013.

[16] Otegi A, Arregi X, Agirre E. Query expansion for ir using knowledge-based relatedness. In: Proceedings of the 5th international joint conference on natural language processing. Thailand; 2011. p. 1467–71.

[17] Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: bringing order to the web. Tech rep. Stanford InfoLab; 1999.

[18] Perez-Iglesias J, Perez-Aguera J, Fresno V, Feinstein Y. Integrating the probabilistic models BM25/BM25F into Lucene. CoRR abs/0911.5046; 2009.

[19] Ramage D, Rafferty AN, Manning CD. Random walks for text semantic similarity. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing. Association for Computational Linguistics; 2009. p. 23–31.

[20] Pakhomov S, Pedersen T, Chute C. Abbreviation and acronym disambiguation in clinical discourse. In: AMIA symposium; 2005. p. 589–93.

[21] Stevenson M, Agirre E, Soroa A. Exploiting domain information for word sense disambiguation of medical documents. JAMIA 2012;19(2):235–40.

[22] Voorhees EM. The trec medical records track. In: Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics. BCB'13. New York (NY, USA): ACM; 2013. p. 239:239–46. <http://doi.acm.org/10.1145/2506583.2506624>.

[23] Voorhees EM, Tong RM. Overview of the TREC 2011 medical records track. In: The tenth text REtrieval conference. National Institute of Standards and Technology, Gaithersburg, MD; 2011.

[24] Winter C, Kristiansen G, Kersting S, Aust JRD, Knösel T, Rümmele P, et al. Google goes cancer: improving outcome prediction for cancer patients by network based ranking of marker genes. PLoS Comput Biol 2012;8(5).

[25] Zhu D, Carterette B. Using multiple external collections for query expansion. In: Medical records track, text retrieval conference (TREC) 2011; 2011.

[26] Zhu D, Carterette B. An adaptive evidence weighting method for medical record search. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. SIGIR '13. New York (NY, USA): ACM; 2013. p. 1025–8. <http://doi.acm.org/10.1145/2484028.2484175>.