Information Retrieval Journal
https://doi.org/10.1007/s10791-018-9344-z

KNOWLEDGE GRAPHS AND SEMANTICS IN TEXT ANALYSIS AND RETRIEVAL

Payoffs and pitfalls in using knowledge-bases for consumer health search

Jimmy[1,2] · **Guido Zuccon**[1] · **Bevan Koopman**[3]

Received: 2 May 2018 / Accepted: 15 October 2018

Abstract
Consumer health search (CHS) is a challenging domain with vocabulary mismatch and considerable domain expertise hampering peoples' ability to formulate effective queries. We posit that using knowledge bases for query reformulation may help alleviate this problem. How to exploit knowledge bases for effective CHS is nontrivial, involving a swathe of key choices and design decisions (many of which are not explored in the literature). Here we rigorously empirically evaluate the impact these different choices have on retrieval effectiveness. A state-of-the-art knowledge-base retrieval model—the Entity Query Feature Expansion model—was used to evaluate these choices, which include: which knowledge base to use (specialised vs. general purpose), how to construct the knowledge base, how to extract entities from queries and map them to entities in the knowledge base, what part of the knowledge base to use for query expansion, and if to augment the knowledge base search process with relevance feedback. While knowledge base retrieval has been proposed as a solution for CHS, this paper delves into the finer details of doing this effectively, highlighting both payoffs and pitfalls. It aims to provide some lessons to others in advancing the state-of-the-art in CHS.

Keywords Knowledge base · Knowledge graph · Query expansion · Consumer health search

✉ Guido Zuccon
g.zuccon@uq.edu.au

Jimmy
jimmy@hdr.qut.edu.au

Bevan Koopman
bevan.koopman@csiro.au

[1] Queensland University of Technology (QUT), Brisbane, QLD, Australia

[2] University of Surabaya (UBAYA), Surabaya, Indonesia

[3] Australian e-Health Research Centre, CSIRO, Canberra, Australia

Published online: 08 November 2018

1 Introduction

A major challenge for users in consumer health search (CHS) is how to effectively represent complex and ambiguous information needs as a query (Zhang 2014; Toms and Latter 2007; Zeng et al. 2002). Studies on query formulation in CHS have shown that consumers struggle to find effective query terms (Zeng et al. 2002), often submitting layman and circumlocutory descriptions of symptoms instead of precise medical terms (Stanton et al. 2014; Zuccon et al. 2015). For example, people search for "skin irregularities" instead of "skin lesions" (the correct medical term for the symptom). They do so using general web search engines, which are commonly preferred over specialised health web sites and services (Fox and Duggan 2013; McDaid and Park 2011). However, previous work has shown that the use of general web search engines for answering these specific health needs leads to poor retrieval effectiveness, incorrect information and possibly low user satisfaction (Zuccon et al. 2015). Different approaches have been proposed to improve CHS, including query suggestion (Zeng et al. 2006), learning-to-rank using syntactic, semantic or readability features (Soldaini and Goharian 2017; Palotti et al. 2016), and query expansion or reformulation (Soldaini et al. 2016; Silva and Lopes 2016; Plovnick and Zeng 2004).

Here we focus on overcoming the problems in CHS by expanding a health query with more effective terms (e.g., less ambiguous, synonyms, etc.). For example, the query "skin tag" can be expanded by adding the term "acrochordon" which is a medical term for skin tag. The term "acrochordon" provides better disambiguation as it effectively represents the original two terms query. Documents containing the term "acrochordon" are more likely to be relevant to the query than documents containing either "skin" or "tag" alone.

A valuable source of medical domain knowledge is contained in carefully curated medical knowledge bases (KBs); for example, the UMLS medical thesaurus.[1] Studies have shown that manually replacing query terms with those from medical knowledge bases has proven effective (Plovnick and Zeng 2004)—but can it be done automatically?

How to effectively utilise the KB to improve retrieval involves a large number of important design decisions. The impact of these different decisions has not been thoroughly and rigorously considered in most previous approaches (Bendersky et al. 2012; Dalton et al. 2014). Thus, in this paper, we also seek to empirically evaluate the impact of a number of different choices in KB retrieval.

Key contributions

- The implementation and evaluation of a state-of-the-art knowledge base retrieval method to consumer health search;
- The impact of implementation choices, including: (i) KB construction; (ii) entity mention extraction; (iii) entity mapping; (iv) source of expansion; (v) use of relevance feedback. We also determine whether the use of a specialised KB is preferred over a general purpose one, or vice versa.

While some of this material is covered in an existing study (Jimmy et al. 2018), this articles includes the following additional contributions:

[1] Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences.

- An extended literature review highlighting key works that have proposed methods to exploit knowledge bases and knowledge graphs for query expansion, both within and outside health search.
- An expanded explanation of the methods by integrating a meaningful example that aids the understanding of the key differences produced by each considered choice in the KB query expansion process.
- The addition of the Consumer Health Vocabulary (CHV) as another knowledge base (Choice 1). CHV provides a mapping between professional medical lingo and consumer expressions (Zeng and Tse 2006; Keselman et al. 2008).
- The extraction of query entity mentions (Choice 2) using Metamap (Aronson and Lang 2010) (a biomedical information extraction system).
- A study of combining expansion terms from all KBs (Wikipedia, UMLS and CHV) when considering the source of expansion for term selection (Choice 4).
- An evaluation of an alternative approach for relevance feedback and pseudo relevance feedback (Choice 5) based on Soldaini et al. (2015)'s work, which filters expansion terms based on their likelihood of being health related.
- An investigation of results generalisability via evaluation using an additional test collection, CLEF eHealth 2015, which used different queries and different websites crawls.
- An analysis of the influence of unjudged documents on retrieval results, including evaluating using the combined relevance assessments from CLEF 2016 and CLEF 2017, and using condensed list approach (Sakai 2007).

The remainder of this paper is structured as follows. Section 2 discusses previous work related to this article. Section 3 describes the query expansion model used and the choices we consider for knowledge base retrieval. Section 4 explains the data collection used in this work. Section 5 details the empirical evaluation performed and the evaluation results. Section 6 analyses and discusses the evaluation results, while Sect. 7 concludes this article. Additionally, "Appendix 1: Statistical significance analysis" section reports the statistical significance analysis for all the results of the experiments discussed in this article, and "Appendix 2: List of abbreviations" section lists the abbreviations used to provide the reader with a quick-to-consult reference.

2 Related work

2.1 Knowledge-base retrieval

Knowledge bases such as Wikipedia and Freebase have been used to automatically improve retrieval effectiveness by augmenting user-issued queries. We start by introducing the method we rely on in this article: the Entity Query Feature Expansion (EQFE) (Dalton et al. 2014) (The actual formulation of the method is detailed in Sect. 3.1). This model performs automated query expansion by linking mentions from the original query to concepts in Wikipedia. Instead of achieving this through a direct mapping (as we later show Bendersky et al. (2012) did), the Entity Query Feature Expansion model labels words in the query and in each document with a set of entity mentions M_Q and M_d (Dalton et al. 2014). Each entity mention is related to KB entities $e \in E$, with different relationship types. Queries are then expanded by including entity aliases, categories, words, and types from their related

Wikipedia articles. The expanded queries are then matched against documents in the corpus using the query likelihood model with Dirichlet smoothing.

We posit that this Entity Query Feature Expansion model is a natural fit for consumer health search. It provides a means of mapping health queries to health entities in a health related (subset of a) KB, be this either a general purpose KB (e.g., Wikipedia) or a domain-specific KB (e.g., UMLS). The initial query can then be expanded based on related entities. In this article, we investigated the use of both a specialised health KB, in line with previous work that expanded queries using, e.g., MeSH or UMLS (Soldaini et al. 2016; Díaz-Galiano et al. 2009; Silva and Lopes 2016), and a general purpose KB, Wikipedia. Our rationale for this latter choice was the observation that consumers tend to submit queries using general terms and that these are covered by Wikipedia entities. However, Wikipedia also covers many of the medical entities found in specialised medical KBs. More importantly, there are links between the general and specialised entities in Wikipedia—links that can be exploited for query expansion. For the same reason, we have further extended the choices we investigated for KB construction by also considering the consumer health vocabulary (CHV), which, like Wikipedia, provides a direct link between professional lingo and consumer expressions (e.g. "myocardial infarction" ⇒ "heart attack"); however, unlike Wikipedia, CHV does this explicitly, rather than implicitly. Thus, we adopted the Entity Query Feature Expansion model for our empirical evaluation, determining if such a KB retrieval approach is effective for CHS.

Other methods for knowledge base retrieval do exist: next we provide a brief account of selected methods used for KB retrieval.

For example, Bendersky et al. (2012) proposed a query formulation approach that links queries to concepts in multiple information sources such as Wikipedia, query logs, and the retrieval corpus itself, using pseudo-relevance feedback. First, they weighted concepts from the query by considering the frequency of each concept found in Google N-grams, MSN Query log , Wikipedia Titles, and the retrieval corpus. Then, a large pool of candidate expansion terms was built for each information source using pseudo-relevance feedback. Candidate expansion terms in the pool were ranked based on their weight as formulated in the first step. The top 100 terms from each pool were then combined and further ranked using a weighted combination of expansion scores. Finally, only the top K terms from the combined pool were used as expansion terms ($K \leq 10$).

Balaneshinkordan and Kotov (2016) empirically investigated the effectiveness in adhoc search tasks of query expansion terms derived from the DBpedia, Freebase and ConceptNet knowledge bases, as well as from the actual document collection. Query expansion terms were derived using information theoretic measures (mutual information) and term associated approaches [term co-occurrence via the Hyperspace Analogous to Language method (Lund and Burgess 1996)]. These were then interpolated with scores from a Dirichlet language model. They found that term associations derived from KBs often provided the highest effectiveness. Compared to Balaneshinkordan and Kotov (2016), we used the more sophisticated EQFE model to select and combine entities to augment the initial user's query. We also took a radically different approach for estimating entity mapping and selection, and further explored more choices available when using KB for query expansion.

Balaneshinkordan and Kotov (2016) found that ConceptNet proved the most effective source of query expansions for general, adhoc tasks. ConceptNet is a KB that represents commonsense knowledge. This is in line with previous work that also found ConceptNet to be a valuable source of expansion terms for adhoc, not domain-specific, searches (Kotov and Zhai 2012). In this article, we have not explored the use of ConceptNet, as terms and associations captured there do not appear to be relevant for CHS. For example, in

ConceptNet, the term "insomnia" is linked to irrelevant, non health-related concepts such as "alternative rock" and "alternative progressive". When links to health-related concepts do they exist, the quality is poor. For example, identified causes of insomnia in ConceptNet are "going to bed", "coffee" and "surfing the net".[2] This is, of course, a very limited account of the causes of insomnia (as identified by the Sleep Foundation).[3]

Xiong and Callan (2015) considered query expansion using Freebase as a KB and, like us, considered the choices involved when setting up systems to do this, including their effectiveness in web search tasks. In contrast, they consider a limited array of choices, including: entity mention extraction (akin to our Choice 2) and selection of expansion terms (we do not have this as the EQFE model is used to determine the expansion terms to be selected). For each of the two choices, they only explored two variants, while we explore many variations for choices in KB retrieval. Specifically, for entity mention extraction they considered either direct (query) keyword match or object frequency from automatic annotations contained in Google's FACC1 annotation set. For selection of expansion terms they considered a pseudo-relevance feedback approach (which somewhat is comparable, in spirit, to our analysis of relevance feedback mechanisms—Choice 5) and a supervised classification approach (SVM).

Liu and Fang (2015) developed a method for entity-based retrieval that represented entity in a latent space and computed retrieval scores by mapping document and query entities to the common entity latent space and then considered the projections of documents and queries in such space. Their approach is alternative to the EQFE method used in this article—a comparison between the latent entity space of Liu and Fang and EQFE in CHS settings is out of the scope of this article; however we intend to direct future work towards this comparison.

The query expansion technique we considered in this work, EQFE, applies entity extraction and analysis to the query expansion stage of the retrieval process. Other techniques, instead, use entities throughout the different stage of retrieval (i.e., in both indexing and retrieval). This is the case, for example, of the concept-based IR model, Explicit Semantic Analysis (Egozi et al. 2011), which relied on entities represented in Wikipedia to identify suitable indexing and retrieval features. A similar approach to concept/entity-based IR had been followed by methods in the medical domain. For example, Zuccon et al. (2012) used the SNOMED-CT terminology to represent medical entities at indexing and retrieval. Their method further exploited subsumption (i.e., parent-child) relationships between entities to derive query expansion terms. While, Koopman et al. (2012) used co-occurrence graphs between entities in the same document for retrieval, also relying on an entity-based indexing and retrieval mechanism. The downside of these methods is that entity indexing is often computationally demanding (e.g., entity extraction and annotation much be run across all document in the corpus) and thus difficult to scale to large web corpora (such as those used in this article).

2.2 Consumer health search (CHS)

One of the major challenges in CHS is the vocabulary mismatch between people's query terms and the terms used in high quality health web resources. One source of high quality

[2] http://conceptnet.io/c/en/insomnia. Last visited 30/04/2018.

[3] https://sleepfoundation.org/insomnia/content/what-causes-insomnia. Last visited 30/04/2018.

health related terms is the Unified Medical Language System (UMLS) (Bodenreider 2004). However, UMLS concepts are rarely mentioned in consumer health queries: Keselman et al. (2008) showed that only 8.1% of 4,928,158 n-grams from consumer queries can be mapped (i.e., exact match) to the UMLS concepts. In this section, we discuss work related to knowledge-base retrieval for CHS.

In constrast, Wikipedia is a crowdsourced, general purpose KB allowing people to promote and describe new concepts or augment existing concepts. While general purposes, Wikipedia contains considerable and detailed health information that has been effectively used in health related information retrieval (Jimmy et al. 2018; Soldaini et al. 2015).

In an earlier study, we evaluated several design choices to instantiate the EQFE model in CHS (Jimmy et al. 2018). These were:

1. Collect pages with medicine infobox[4] type[5] (e.g., "abortion method", "alternative medicine", "pandemic");
2. Collect pages with health infobox type or with links to medical terminologies such as UMLS, Disease DB and ICD in the health infobox;
3. Collect pages that had a least one UMLS entity mention in their title. Entity extraction was done using QuickUMLS (Soldaini and Goharian 2016).

Previously, Soldaini et al. (2015) utilised Wikipedia to select health related terms from clinical case reports. First, they built a health related Wikipedia KB by collecting pages that contained infobox with links to medical terminologies and a non-health related Wikipedia KB that contained the other pages. Then, they calculated the probability of a term being health related by computing the ratio between the probability of the term being found in the health KB and that of the term being found in the non-health KB. We employed a similar method to limit the terms selected by relevance feedback (RF) processes (either explicit or pseudo RF) (see Sect. 3.2.5).

The probability of a term being health related is also an effective method to select expansion terms for CHS (Soldaini et al. 2016). Here medical synonyms were extracted by mapping query terms to 3 medical KBs (Behavioral, MedSyn, or DBpedia). Then, a synonym with the highest probability of being health related was added to the original query. Finally, a supervised classifier was used to select the most likely synonym for each query. In our study, we further explored features of KB (beyond synonyms) to improve the effectiveness of CHS queries.

In contrast with Wikipedia, the UMLS is a medical specific knowledge base that contains medical concepts and relationships among concepts (Bodenreider 2004). Its latest 2017 version (i.e., 2017AB) contains approximately 3.64 million concepts that are compiled from 201 biomedical vocabularies in various languages. Each UMLS concept is grouped into one or more semantic types (out of 133 semantic types in total). As the UMLS is compiled from biomedical vocabularies, it contains many semantic types that are not relevant to CHS such as amino acid sequence, cell function, embryonic structure, etc. For this reason, Soldaini et al. (2016) and Limsopatham et al. (2013) decided to include only concepts from 16 semantic types that were considered as related to the four aspects of

[4] A Wikipedia Infobox is used to summarise important aspects of an entity and its relation with other articles.

[5] http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes#Health_and_fitness.

medical decision criteria: symptom, diagnostic test, diagnosis, and treatment. In our experiments using the UMLS, we follow the same practice.

Using UMLS for CHS still results in vocabulary mismatch between people queries and the medical term in the UMLS (Keselman et al. 2008). To overcome this, the Consumer Health Vocabulary (CHV) (Zeng and Tse 2006; Keselman et al. 2008) was built; this open access resource provides a mapping between consumer health terms and UMLS concepts.

This mapping is constructed by extracting n-grams from MedlinePlus queries and various health-focused bulletin boards; then, automatically mapping these n-grams to UMLS via exact match comparison. Any un-mapped n-grams are then manually mapped to the UMLS (Keselman et al. 2008). From 2007, the CHV is available as part of the UMLS entries with "CHV" as source (i.e., SAB).

Both UMLS and Wikipedia have been used as learning to rank features (LtR) for CHS (Soldaini and Goharian 2017). The results showed that using Wikipedia average *idf* and *tf* in health pages were the first and third best LtR features, respectively. Using UMLS, the number of matching UMLS concepts in document, the number of "sign or symptom" concepts found in a document, and the number of "injury or poisoning" concepts found in document were the second, fifth, and seventh best LtR features, respectively. The best LtR system from Soldaini and Goharian (2017) beat a baseline system by 26.6% on the CLEF2016 dataset (nDCG@10: 0.305 vs nDCG@10: 0.241). This is the same dataset used in this article; thus, we used the results of their study as a benchmark.

In this study, we posit that Wikipedia, UMLS, and CHV have the potential to improve the consumer health search. We evaluated the effectiveness of various CHS design choices using these three KBs.

3 Methodology

3.1 Expansion model

We implemented the Entity Query Feature Expansion (EQFE) model for retrieval on the Wikipedia, UMLS, and CHV as the KB. The EQFE model aims to enrich a query with features from KB entities that are linked to the query. For the Wikipedia KB, a single entity is represented by a single Wikipedia page (the page title identifies the entity). Beyond titles, Wikipedia also contains many page features useful in a retrieval scenario: entity title (E), categories (C), links (L), aliases (A), and body (B). As for the UMLS and CHV KBs, a single entity is represented by the most frequently used terms for a single concept unique identifier (CUI). Features of a UMLS and CHV KB entity are aliases (A), body (B), parent concepts (P), and related concepts (R). Figure 1 shows the features we used for mapping the queries to entities in the KB and as the source of expansion terms. We formally define the query expansion model as:

$$\hat{\vartheta}_q = \sum_{M} \sum_{f} \lambda_f \vartheta_{f(EM,SE)} \tag{1}$$

where M are the entity mentions and contain uni-, bi-, and tri-gram generated from the query; f is a function used to extract the expansion terms. $\lambda_f \epsilon (0, 1)$ is a weighting factor. $\vartheta_{f(EM,SE)}$ is a function to map entity mention M to the KB features EM (e.g., "Title", "Aliases", "Links", "Body", etc.) and extract expansion terms from source of expansion SE (e.g.,"Title", "Aliases", etc.).

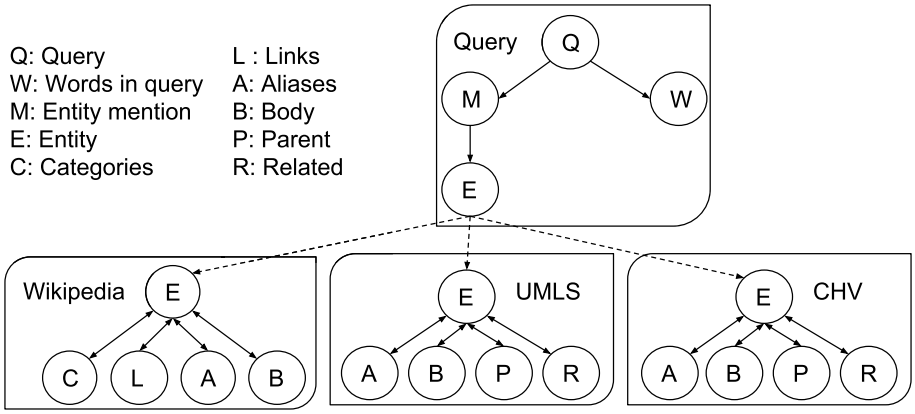

Fig. 1 Summary of expansion sources

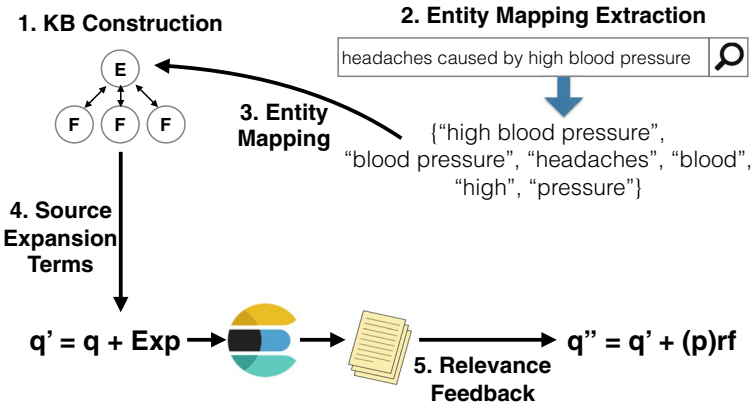

Fig. 2 The EQFE pipeline we considered in this article when instantiating this model. In this model, q is the original query, q' is an expanded query, Exp is the expansion terms, and q" is a query expanded with (pseudo-) relevance feedback (p(rf)), after the original query was augmented using query expansion

3.2 Choices in knowledge base retrieval

This section describes the choices that we considered for each component of the EQFE pipeline (Fig. 2). To select the expansion terms, first, we constructed a number of knowledge bases (KBs). Each KB contains features such as title, aliases, etc. Second, we extracted entities from the original queries. Third, we mapped the query entities to entities in each KB by exact matching each query entity to every KB's features. Fourth, we sourced expansion terms from the mapped KB entities' features. Finally, fifth, we performed relevance feedback with the aim to further improve the already expanded queries. The remainder of this section will describe the choices in details.

3.2.1 Choice 1: knowledge base construction

We investigated which entities should form the basis of our KB. The CHS focus meant that health-related entities were needed. For Wikipedia KB, we considered four Wikipedia Construction (WC) choices for collecting health related pages:

WC-All: all wikipedia pages;
WC-Type: pages with Medicine infobox[6] type[7] (e.g., "abortion method", "alternative medicine", "pandemic");
WC-TypeLinks: pages with Medicine infobox type and pages with infobox containing links to medical terminologies such as Mesh, UMLS, SNOMED CT, ICD;
WC-UMLS: pages with title matching an UMLS entity.

The last method used QuickUMLS (Soldaini and Goharian 2016) to map Wikipedia page titles to the UMLS: if the mapping was successful, we included the Wikipedia entity (page) in the KB.

For UMLS and CHV KBs, we considered the following UMLS Construction (UC) and CHV Construction (CC) choices:

UC/CC-All: all entities;
UC/CC-Med: entities related to four key aspects of medical decision criteria (i.e., symptoms, diagnostic test, diagnoses, and treatments) as used in (Limsopatham et al. 2013; Soldaini et al. 2016).

For these choices, we included all English and non-obsolete terms.

3.2.2 Choice 2: entity mention extraction

Entity mention extraction is the process of identifying spans of text from the query that could map to some entity, while it does not consider which exact entity a span is mapped to (this is detailed in the next section). We considered four possible Mention Extraction (ME) choices to extract entity mentions (see Fig. 3):

ME-All: include all uni-, bi- and tri-grams of the query (*default choice*);
ME-CHV: include only those uni-, bi- and tri-grams of the query that matched entities in the Consumer Health Vocabulary (CHV) (Keselman et al. 2006);[8]
ME-UMLS: include only those uni-, bi- and tri-grams of the query that matched entities in the UMLS (via QuickUMLS);
ME-MetaMap: include only those uni-, bi- and tri-grams of the query that matched health entities via MetaMap (Aronson and Lang 2010).

[6] A Wikipedia Infobox is used to summarise important aspects of an entity and its relation with other articles.
[7] http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes#Health_and_fitness.
[8] Only complete string matches were considered.

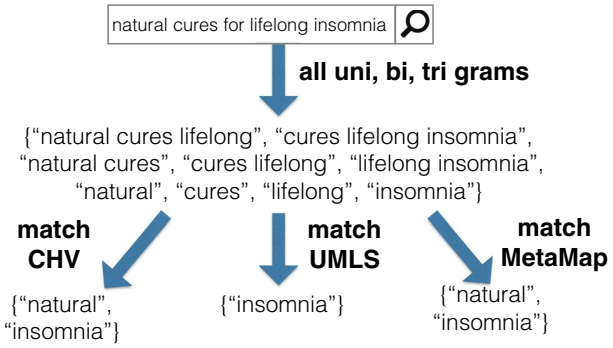

Fig. 3 Extracting entity mentions from the query "natural cures for lifelong insomnia": the influence of different choices for entity extraction (Choice 2)

These choices were used for all KBs. For ME-CHV, we used the CHV version included in the UMLS version 2017AB [while in our previous work we used CHV version 20110204 (Jimmy et al. 2018)].

3.2.3 Choice 3: entity mapping

We investigated how the entity mentions from the previous section were mapped to entities in the KB. An entity mention was mapped to an entity if an exact match was found between the mention and the entity. As shown in Fig. 1, the Wikipedia entity can be represented according to five different features. The Wikipedia Entity Mapping (WEM) choices considered were:

WEM-Title: titles;
WEM-Aliases: aliases;
WEM-Links: links;
WEM-Body: the entire bodies of the Wikipedia pages;
WEM-Cat: categories;
WEM-All: all the previous sources (*default choice*).

For UMLS and CHV KBs, the UMLS Entity Mapping (UEM) and CHV Entity Mapping (CEM) choices considered were:

UEM/CEM-Title: titles;
UEM/CEM-Aliases: aliases;
UEM/CEM-Body: the entire UMLS concept description;
UEM/CEM-Parent: parents;
UEM/CEM-Related: related entities;
UEM/CEM-All: all the previous sources (*default choice*);
UEM/CEM-QuickUmls: use QuickUMLS to obtain entity mappings.

Table 1 shows the mappings to the Aliases feature of each KB for the query "abdominal pain, vomiting, pain near belly button, duplicated ureter".

Table 1 Choice 3: Mapped entities for query id 122006: "abdominal pain, vomiting, pain near belly button, duplicated ureter" are mapped to the Aliases feature of each KB

Wikipedia TypeLinks	UMLS All	CHV Med
"abdominal pain", "navel", "abdomen"	"umbilicus", "double ureter", "muscle belly", "duplication", "procedures on ureter", "vomiting adverse event"	"umbilicus", "muscle belly", "1/3 meter"

3.2.4 Choice 4: source of expansion

We investigated which sources in the KB were used to draw candidate terms for query expansion. We explored three Source of Expansion (SE) choices:

SE-Title: titles associated with the entities;
SE-Aliases: aliases associated with the entities;
SE-All: both titles and aliases (*default choice*).

While other information sources could be used (for example, those used for entity mapping), preliminary experiments showed that only these three choices produced meaningful results. These choices were used for all KBs (Wikipedia, UMLS, and CHV). An example of the different outputs obtained by each variant for this choice is shown in Table 2.

3.2.5 Choice 5: relevance feedback

The unique challenges of CHS make explicit relevance feedback (RF, i.e., where feedback comes from the user) a worthwhile consideration for improving retrieval effectiveness. The question that follows is what gains are possible if the user was providing explicit feedback? To answer this we apply RF by using the actual relevance labels (qrels) to simulate an accurate user selecting relevant documents. Comparison is made to a non-RF baseline to determine the effective gain from explicit RF. In this study, we investigated the use of relevance feedback (both explicit relevance feedback (RF) and Pseudo Relevance Feedback (PRF)) as used in Jimmy et al. (2018).

We performed RF by extracting the 10 most important health related words (based on tf.idf scores) from each of the top three relevant documents (relevance label greater than 0) thus resulting in a maximum of thirty expansion terms. PRF was performed by extracting the 10 most important health related words from the top three ranked documents (regardless of their true relevance label). A term was considered as health related if it exactly matched a title or an alias of an entity in the target KB: either Wikipedia (WC-TypeLinks) or UMLS (UC-All).

In addition, in this study we also considered the relevance feedback approach proposed by Soldaini et al. (2015). We refer to this approach as RF Health Terms (RFHT) and PRF Health Terms (PRFHT), as they filtered the candidate relevance feedback terms based on the probability of the term being health related, based on likelihoods computed from Wikipedia (see Sect. 2.2).

In PRFHT, all terms in the top k results with high probability of being health-related are extracted and used for query expansion. This probability is calculated as:

Table 2 Choice 4: Expansion terms selected for each KB when considering different variants for the choice source of expansion. For this example, the initial query was id 103004: "headaches caused by too much blood or "high blood pressure""

Source	Wikipedia	UMLS	CHV
Title	Hypertension	Hypertension,finding, peripheral, abnormally	Hypertension, arterial
Aliases	Signs, hypertension, arterial, chronic, disease, hyperpiesia, hyperpieses, awareness, epidemiology, economics, disorders, accelerated, prognosis, symptoms, diagnosis, hypertention, residual, raised, bp, prevention, refractory, adrenal, elevated, hyper, tension, classification, increased, rebound, taking, presure, venous, systolic, bloodpressure, measuring, msdbp, human, diastolic, leg, arm, index, pulmonary, nibp, invasive, regulation atrial, determination, normotensive	Hypertension, systemic, vascular, disease, disorder, hyperpiesis, htn, arterial, hbp, elevated, cardio pulm, tension, ht, bp, hyperpiesia, finding, increased, i10, i15, degeneration, diagnosis, result, peripheral, substance reticuloendothelial, whole, abnormally	Systemic, hypertension, vascular, disease, disorder, hyperpiesis, htn, arterial, hbp, elevated, cardio, pulm, tension, ht, bp, hyperpiesia, finding, increased, i10, i15, degeneration, diagnosis

$$OR(t_j) = \frac{Pr\{P \text{ is health related } | t_j \epsilon P\}}{Pr\{P \text{ is not health related } | t_j \epsilon P\}} \tag{2}$$

where P is a Wikipedia page and term t_j is included in a query if $OR(t_j) \geq \delta$. In our experiments, we calculated the probabilities of a Wikipedia page P being health related and being not-health related as:

$$Pr\{P \text{ is health related } | t_j \epsilon P\} = \frac{|P \epsilon D_h : t_j \epsilon P|}{|D_h|} \tag{3}$$

$$Pr\{P \text{ is not health related } | t_j \epsilon P\} = \frac{|P \epsilon D_{nh} : t_j \epsilon P|}{|D_{nh}|} \tag{4}$$

where D_h is a collection of Wikipedia pages with health infobox and links to medical terminologies (i.e., WC-TypeLinks) and D_{nh} contains Wikipedia pages that are not included in D_h. Using the English subset of Wikipedia crawled on the 1/12/2016, we found that $|D_h| = 13{,}135$ and $|D_{nh}| = 9{,}182{,}304$.

While Soldaini et al. (2015) suggested that the optimal value for δ is 2, in preliminary experiments we found that $\delta = 2$ is too low, as many non-health terms scored $\delta \geq 2$; in this study, instead, we used $\delta = 4$ as it was a better fit. This difference was likely due to a different Wikipedia dump being used: ours was substantially larger than that reported by Soldaini et al. Further, to prevent query drift, we further limited the number of expansion terms added for PRFHT to 20.

Once terms are filtered to retain only terms estimated to be health related, the j-th health term in document D_i is weighted according to:

$$b_j = log_{10}(10 + w_j) \tag{5}$$

where:

$$w_j = \alpha \,.\, I_q(t_j) \,.\, tf_j + \left(\frac{\beta}{k}\right) \,.\, \sum_{i=1}^{k} I_{D_i}(t_j).idf_j \tag{6}$$

Following the work of Soldaini et al. (2015), we fixed $k = 10$, $\alpha = 2.0$ and $\beta = 0.75$. In Eq. 6, $I_q(t_j) = 1$ if $t_j \in Q$, and 0 if otherwise. $I_{D_i}(t_j) = 1$ if $t_j \in D_i$, and 0 if otherwise.

For the explicit relevance feedback (RFHT), we modified the above PRFHT approach to only extract terms from the top k explicitly relevant documents. Unlike the PRFHT, for RFHT, we did not limit the number of expansion terms added: all expansion terms with $\delta \geq 4$ were added to the original query.

4 Data collection

To investigate the influence choices in KB retrieval have on query expansion for the CHS task, we empirically evaluated methods using the CLEF 2016 eHealth (Zuccon et al. 2016). This collection comprises 300 query topics originating from health consumers seeking health advice online. Documents are taken from Clueweb12b-13. The collection was indexed using Elasticsearch 5.1.1, with stopping and stemming. A simple baseline was implemented using BM25F with $b = 0.75$ and $k1 = 1.2$. BM25F allows specifying boosting factors for matches occurring in different fields of the indexed web page. We considered only the title field and the body field, with boost factors 1 and 3, respectively. These were found to be the optimal weights for BM25F for this test collection in previous work (Jimmy et al. 2016). This is a strong baseline as it outperforms most runs submitted to CLEF 2016.

For constructing the Wikipedia KB, we considered candidate pages from the English subset of Wikipedia (dump 1/12/2016), limited to current revisions only and without talk or user pages. Of the 17 million entries, we filtered out pages that were redirects; this resulted in a Wikipedia corpus of 9,195,439 pages (i.e., WC-All). These candidate pages were then processed according to the choices available for KB construction (Sect. 3.2.1). The total number of pages included in WC-Type is 9562 pages, in WC-TypeLinks is 13,135 pages, and in WC-UMLS is 1,112,206 pages. Selected pages to be included in the KB were also indexed using Elasticsearch 5.1.1 with field based indexing, to support the use of different fields as the source of query expansion terms (Sect. 3.2.4). For all Wikipedia KBs, we indexed the following fields: title (text node of element node `<title>`), links (outbound links to other Wikipedia pages), categories (as defined in `[[Category: category name]]`), types (types of all infoboxes in a page), aliases (text node of element node `<title>` from the page's redirects), and body (text node of element node `<text>`).

For constructing the UMLS KB, we indexed non obsolete English terms (i.e., UC-All) with the following fields: title (the most frequently used term for a CUI), aliases (for all other terms used for the CUI), body (the description of a CUI), parent (title of UMLS entities with relationship type PAR), related (title of UMLS entities with relationship type RQ and RL). Similar to the Wikipedia KB, we processed these UMLS terms according to choices in constructing UMLS KB as described in Sect. 3.2.1 and obtained 3,057,234 terms in UC-All and 1,344,941 UMLS terms in UC-Med.

The CHV KB was constructed by selecting UMLS KB entries with the UMLS SAB field equal to "CHV". The CHV KB index structure was identical to the UMLS KB. For the CHV based KB, we obtained 56,350 terms in CHV-All and 34,514 terms in CHV-Med.

5 Empirical evaluation

Results were evaluated using nDCG@10, RBP@10 (persistence 0.5, depth 10, reporting also residuals (Res.)), in line with the CLEF 2016 collection, as users in the CHS task tend to primarily examine the first few search results. Additionally, bpref was used as a first attempt to reduce the influence of unjudged documents on evaluation (expanded queries retrieved many more unjudged documents than the baseline). For brevity, a full account of statistical significant differences (pairwise t-test with Bonferroni adjustment and $\alpha < 0.05$) between results is reported in "Appendix 1: Statistical significance analysis" section. Furthermore the average number of terms added in the expanded query ($\overline{|exp|}$) and the number of expanded queries, queries with a gain for RBP@10 and a loss for RBP@10 were recorded as a triplet $<e, g, l>$.

For each choice, we empirically evaluated the influence the choice had on retrieval effectiveness by examining each choice sequentially. We did this for all KBs, and drew conclusions about which KB best supports CHS at the end. For each choice, we fixed the best setting and use this best setting for the subsequent choice. We determined the best setting firstly based on results (i.e., nDCG@10, bpref, RBP@10) for all queries set. If no method was clearly best for this set, then we checked results from the high coverage queries set. Lastly, if results from the high coverage queries set were unable to clearly determine which method was best, then we selected the setting with the highest RBP@10 for all queries set as the best setting (RBP@10 was a primary measure for CLEF 2016). The complete set of results is provided in an online appendix at http://ielab.io/kb-chs, along with all run and software source code used.

5.1 Choice 1: knowledge base construction

The effect on retrieval of choices in KB construction is reported in Table 3 (top); results are averaged over all 300 queries in the CLEF 2016 collection.

The results for the Wikipedia KB showed that choice WC-TypeLinks (i.e., pages with health infobox type and links to health terms) lead to the highest effectiveness across all measures. For the UMLS KB, UC-All performed the highest effectiveness on all measures. Lastly, for CHV KB, CC-Med performed the highest across all measures. Nevertheless, the baseline performed considerably better than any KB retrieval method.

When further analysing the results, we found that, for a large number of queries, the KB retrieval methods ranked many unjudged documents amongst the top 10; while the baseline had a much lower rate of unjudged documents amongst the top 10. Figure 4 reported the distribution of unjudged documents for each of the configurations considered. This is clearly influencing the results, as demonstrated by the large values of RBP residuals associated with the KB retrieval methods in Table 3 top (compared to the residual of the baseline). Interestingly, if all unjudged documents turned out to be relevant, the RBP@10 of the

Table 3 Influence of choices in KB construction for CLEF2016 (Choice 1). Statistical significance differences reported in Table 16

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	⟨e,g,l⟩
The all queries set						
Baseline	.2465	.1798	.3263	.0399		
WC-All	.1010	.1512	.1269	.6444	26.28	299, 44, 165
WC-Type	.0982	.1491	.1329	.7006	38.95	299, 59, 157
WC-TypeLinks	**.1146**	**.1547**	**.1532**	.6361	43.22	300, 66, 157
WC-UMLS	.1090	.1475	.1439	.6342	21.17	299, 54, 163
UC-All	**.1256**	**.1653**	**.1626**	.5976	29.27	299, 63, 164
UC-Med	.1189	.1610	.1453	.6004	26.88	298, 54, 168
CC-All	.1108	.1540	.1464	.6251	42.86	299, 52, 171
CC-Med	**.1384**	**.1607**	**.1877**	.5780	36.51	299, 68, 155
The high coverage queries set						
Baseline	.4135	.4684	.4634	.0010		
WC-All	**.5104**	**.4676**	.5364	.1261	19.67	12, 8, 2
WC-Type	.4554	.4105	.4623	.1039	25.75	12, 6, 4
WC-TypeLinks	.4556	.4234	.4534	.1215	24.67	12, 5, 5
WC-UMLS	.4417	.4150	**.6732**	.1678	16.25	12, 9, 3
UC-All	**.3920**	**.3708**	**.5536**	.0150	31.17	12, 8, 3
UC-Med	.2944	.3174	.3932	.0365	32.25	12, 7, 4
CC-All	**.2522**	**.3242**	**.3620**	.1984	41.83	12, 4, 7
CC-Med	.2339	.2956	.3372	.1935	40.25	12, 4, 7

Bold indicates the highest effectiveness achieved for each KB

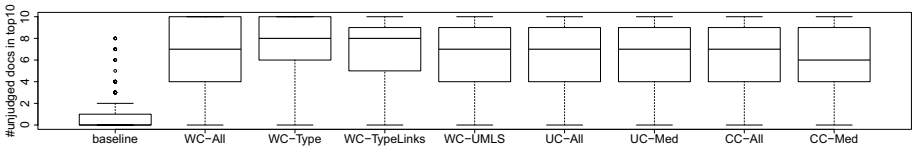

Fig. 4 Unjudged documents among the top 10 retrieved by runs in Table 3 (top)

KB retrieval methods would prove largely superior than that of the baseline (compare the residuals).

We then considered a subset of queries for which, on average across all runs considered for a specific choice, there were a maximum of 2 unjudged documents out of the first 10. This threshold was determined by analysing the number of unjudged documents for the baseline (the baseline does not change, irrespective of the choices), so that the threshold corresponded to 1.5 times the interquartile range above the third quartile (the upper whisker of the box-plot). Note that this produced a different subset of queries for each of the considered choices; however, the subsets had the same average "coverage" with respect to the relevance assessments. We referred to these subsets as the *high coverage queries set*. We instead refer to the set containing all the queries as the *all queries set*. This subset included 12 queries for choice 1 (Table 3, bottom). Results showed reduced residuals and reduced gaps between KB retrieval methods and the baselines; this affected trends in effectiveness across the considered choices for the Wikipedia KB.

Results from Wikipedia KB showed that, for the all queries set, the WC-TypeLinks setting performed best in all three measures. Therefore, although results from the high coverage queries set showed different results, we decided that constructing the Wikipedia KB using the WC-TypeLinks setting was the best option.

Trends in effectiveness for UMLS KB showed that UC-All consistently performed best in both the all queries set and the high coverage queries set. Therefore, we selected UC-All for the following analyses. Lastly, for CHV KB, we found that CC-MED performed best for all queries for all three measures.Thus, we selected CC-Med as the best setting for CHV KB.

Interestingly, the KB constructed with the UC-All choice (that contains many concepts unrelated to the health domain, such as `C0030561: Paris, France`) performed better than the one constructed with the UC-Med choice (that intuitively would contain more health concepts). As noted in Sect. 4, however, the number of concepts in UC-Med are less than half than those of UC-All. It is likely that there exists a better way to filter out non-health related concepts from the UMLS. Based on this, an avenue for future work is an effective method for selecting the subset of UMLS relevant to CHS queries (i.e., improving the construction of the KB based on the UC-Med setting).

5.2 Choice 2: entity mentions extraction

Table 4 (top: 300 queries and bottom: 19 high coverage queries) reports the results obtained when comparing choices for entity mention extraction. For Wikipedia KB, results from the all queries set (Table 4 top) showed no choice was clearly best. Then, we looked at results from the high coverage queries set. Results from the high coverage queries set showed that the WME-CHV setting performed best for all measures. Therefore, we selected WME-CHV as the best setting for Wikipedia KB and used this settings in the following analyses.

For UMLS KB, we found that UME-UMLS performed best for the all queries set for all three measures. Thus, we selected UME-UMLS as the best setting for UMLS KB.

Lastly, for CHV KB, both the all queries set and the high coverage queries set showed no choice was clearly best. Therefore, we selected CME-CHV as the setting for CHV KB as it performed best for RBP@10 in the all queries set.

5.3 Choice 3: entity mapping

Table 5 (top: 300 queries and bottom: 18 high coverage queries) reports the results obtained when comparing choices for entity mapping. For all KBs, mapping entities to Aliases (WEM-Aliases, UEM-Aliases, and CEM-Aliases) clearly outperformed the other approaches (all queries). Results for the high coverage queries showed mixed results. Thus, we selected WEM-Aliases, UEM-Aliases, and CEM-Aliases for the subsequent analyses.

5.4 Choice 4: source of expansion

Table 6 (top: 300 queries and bottom: 129 high coverage queries) reports the results obtained when comparing sources of query expansion. Results clearly showed that selecting titles as source of expansion (WSE-Title, USE-Title and CSE-Title) was the

Table 4 Influence of choices in entity mention extraction (Choice 2). Statistical significance differences reported in Table 17

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	⟨e,g,l⟩
The all queries set						
Baseline	.2465	.1798	.3263	.0399		
WME-All	.1146	.1547	.1532	.6361	43.22	300, 66, 157
WME-CHV	**.1264**	.1573	.1723	.6074	37.34	293, 66, 156
WME-UMLS	.1252	**.1638**	**.1739**	.5901	33.17	297, 67, 154
WME-Metamap	.1166	.1533	.1538	.6277	37.73	296, 59, 162
UME-All	.1256	.1653	.1626	.5976	29.27	299, 63, 164
UME-CHV	.1250	.1659	.1613	.5986	27.31	298, 60, 166
UME-UMLS	**.1304**	**.1702**	**.1728**	.5521	23.79	296, 66, 159
UME-Metamap	.1229	.1633	.1561	.6067	26.79	297, 60, 164
CME-All	.1384	.1607	.1877	.5780	36.51	299, 68, 155
CME-CHV	**.1454**	.1629	**.1953**	.5636	34.13	298, 71, 154
CME-UMLS	.1452	**.1692**	.1941	.5398	30.69	297, 76, 152
CME-Metamap	.1367	.1580	.1839	.5790	34.01	299, 68, 156
The high coverage queries set						
Baseline	.4214	.4036	.4798	.0011		
WME-All	.4401	.3750	.4951	.1324	30.16	19, 7, 9
WME-CHV	**.4593**	**.3807**	**.5268**	.1128	31.44	16, 8, 7
WME-UMLS	.4217	.3658	.5005	.0804	21.11	19, 6, 8
WME-Metamap	.4543	.3797	.5126	.1062	32.12	16, 7, 8
UME-All	.4240	.3906	**.5124**	.1328	33.68	19, 12, 6
UME-CHV	**.4286**	**.3927**	.5117	.1342	32.79	19, 12, 6
UME-UMLS	.4124	.3749	.4929	.1479	29.11	19, 11, 7
UME-Metamap	.4239	.3910	.5123	.1340	33.32	19, 12, 6
CME-All	.3401	.3350	.4485	.1792	41.74	19, 9, 9
CME-CHV	**.3444**	.3361	.4532	.1708	41.11	19, 9, 9
CME-UMLS	.3353	**.3443**	.4369	.1773	36.63	19, 10, 8
CME-Metamap	.3434	.3349	**.4559**	.1792	41.58	19, 10, 8

Bold indicates the highest effectiveness achieved for each KB

most effective choice compared to other choices for both Wikipedia KB and UMLS KB. Therefore, we selected WSE-Title, USE-Title, and CSE-Title as the best settings for each corresponding KB.

Then, we investigated the merit of combining expansion terms from the best setting of each KB; e.g., expansion terms for the WikiChv were generated by combining expansion terms from the WSE-Title and CSE-Title settings. In total, we generated four possible combinations: WikiUmlsChv, WikiUmls, WikiChv, and UmlsChv. Results for both the all queries set and the high coverage queries set showed that no choice was clearly best. We then selected WikiChv as the best setting as it returned the highest RBP@10 for the all queries set.

Table 5 Influence of choices in entity mapping (Choice 3). Statistical significance differences reported in Tables 18 and 19

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	⟨e,g,l⟩
The all queries set						
Baseline	.2465	.1798	.3263	.0399	1.00	
WEM-Title	.1537	.1614	.1930	.3752	25.42	177, 33, 107
WEM-Aliases	**.1999**	**.1691**	**.2692**	.2428	16.83	115, 32, 61
WEM-Links	.1390	.1452	.1943	.3371	22.88	170, 27, 105
WEM-Body	.1375	.1609	.1821	.4195	69.38	226, 47, 132
WEM-Cat	.1785	.1630	.2323	.2683	24.95	108, 23, 70
WEM-All	.1264	.1573	.1723	.6074	37.34	293, 66, 156
UEM-Title	.1540	.1795	.1768	.4938	15.17	272, 48, 156
UEM-Aliases	**.1775**	**.1877**	**.2453**	.3525	9.26	247, 72, 115
UEM-Body	.0789	.1453	.0983	.6487	86.74	289, 40, 177
UEM-Parent	.1370	.1484	.1807	.4982	25.10	243, 43, 134
UEM-Related	.1531	.1684	.1989	.4808	29.43	260, 63, 139
UEM-All	.1304	.1702	.1728	.5521	23.79	296, 66, 159
UEM-QuickUmls	.1355	.1792	.1563	.5550	30.37	297, 65, 162
CEM-Title	.1704	.1796	.2023	.3812	10.93	211, 36, 127
CEM-Aliases	**.2142**	**.1858**	**.2936**	.2903	11.81	185, 63, 77
CEM-Body	.1196	.1521	.1487	.6131	43.76	271, 45, 152
CEM-Parent	.1252	.1483	.1712	.5235	52.48	252, 46, 141
CEM-Related	.1669	.1762	.2370	.4703	29.34	262, 85, 122
CEM-All	.1454	.1629	.1953	.5636	34.13	298, 71, 154
CEM-QuickUmls	.1543	.1791	.1788	.5337	22.61	279, 67, 149
The high coverage queries set						
Baseline	.3650	.3820	.4074	.0010		
WEM-Title	.3943	**.3892**	.4164	.0676	16.88	8, 4, 4
WEM-Aliases	.3398	.3803	.3883	.0491	12.78	9, 4, 5
WEM-Links	.2593	.2856	.3074	.1402	27.50	12, 3, 9
WEM-Body	.2521	.2870	.2909	.0447	88.81	16, 5, 11
WEM-Cat	.3529	.3771	.3952	.0010	3.50	2, 0, 2
WEM-All	**.4184**	.3656	**.4533**	.2176	23.40	15, 7, 6
UEM-Title	**.4494**	.3669	**.5181**	.1445	10.07	14, 7, 6
UEM-Aliases	.3635	**.3800**	.4417	.0480	16.13	15, 8, 6
UEM-Body	.1934	.2468	.1512	.3416	60.94	16, 1, 13
UEM-Parent	.3394	.3360	.4021	.1194	27.25	12, 8, 4
UEM-Related	.3848	.3724	.4710	.0235	25.00	13, 9, 3
UEM-All	.3542	.3342	.4723	.1052	27.75	16, 9, 6
UEM-QuickUmls	.3766	.3529	.4447	.1094	27.56	16, 8, 7
CEM-Title	.3592	**.3945**	.3746	.0783	6.60	5, 0, 5
CEM-Aliases	**.4002**	.3554	**.5018**	.0772	18.60	15, 10, 3
CEM-Body	.3730	.3436	.4292	.1811	33.67	12, 5, 6
CEM-Parent	.2339	.3302	.2642	.1718	36.67	15, 4, 10
CEM-Related	.3682	.3613	.4885	.0704	23.62	16, 11, 4
CEM-All	.2664	.2998	.3528	.2514	37.41	17, 6, 9
CEM-QuickUmls	.3809	.3606	.4661	.0956	23.25	16, 8, 7

Table 5 (continued)

Bold indicates the highest effectiveness achieved for each KB

Table 6 Influence of choices in source of expansion (Choice 4). Statistical significance differences reported in Table 20

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	⟨e,g,l⟩
The all queries set						
Baseline	.2465	.1798	.3263	.0399	1.00	
WSE-Title	**.2430**	**.1843**	**.3231**	.0824	1.37	76, 27, 32
WSE-Aliases	.1991	.1689	.2688	.2412	16.60	115, 31, 62
WSE-All	.1999	.1691	.2692	.2428	16.83	115, 32, 61
USE-Title	**.2137**	**.1917**	**.2961**	.2115	2.60	217, 67, 94
USE-Aliases	.1910	.1892	.2599	.3163	8.96	228, 71, 99
USE-All	.1775	.1877	.2453	.3525	9.26	247, 72, 115
CSE-Title	**.2433**	**.1929**	**.3283**	.1236	1.77	155, 59, 58
CSE-Aliases	.2143	.1858	.2941	.2869	11.77	185, 63, 77
CSE-All	.2142	.1858	.2936	.2903	11.81	185, 63, 77
WikiUmlsChv	.2272	**.1972**	.3187	.2290	3.17	246, 83, 101
WikiUmls	.2187	.1945	.3033	.2247	2.79	232, 73, 100
WikiChv	**.2441**	.1954	**.3300**	.1409	1.98	181, 69, 70
UmlsChv	.2222	.1941	.3106	.2232	3.00	232, 79, 94
The high coverage queries set						
Baseline	.3025	.2260	.3851	.0141		
WSE-Title	**.3054**	.2298	**.3877**	.0265	1.26	23, 11, 9
WSE-Aliases	.2788	**.2305**	.3618	.1177	11.98	41, 19, 19
WSE-All	.2806	.2304	.3619	.1178	12.15	41, 20, 18
USE-Title	**.3109**	**.2408**	**.4073**	.0311	2.09	78, 38, 20
USE-Aliases	.2766	.2326	.3663	.1005	8.55	91, 32, 34
USE-All	.2755	.2317	.3638	.1047	9.08	93, 35, 34
CSE-Title	**.3231**	**.2422**	**.4273**	.0182	1.75	48, 28, 12
CSE-Aliases	.3007	.2325	.3880	.0787	11.67	64, 26, 22
CSE-All	.3007	.2325	.3880	.0787	11.67	64, 26, 22
WikiUmlsChv	.3190	.2420	.4266	.0408	2.48	89, 4, 22
WikiUmls	.3093	.2413	.4031	.0421	2.11	87, 40, 25
WikiChv	**.3223**	**.2427**	.4232	.0296	1.79	58, 31, 17
UmlsChv	.3198	.2414	**.4307**	.0294	2.51	80, 44, 17

Bold indicates the highest effectiveness achieved for each KB

5.5 Choice 5: relevance feedback

Table 7 (top 300 queries and bottom: 76 high coverage queries) reports the results obtained with and without relevance feedback. For the all queries set, results for All KBs showed that the addition of relevant feedback filtered based on the likelihood of being health related (RFHT) performed the best across all measures. On the contrary, the addition of

pseudo relevant feedback hurted the performance for all KBs (with the exception of baselinePRFHT and CSE-TitlePRFHT that had a better bpref than the baseline and CSE-Title without the pseudo relevance feedback).

Results from the high coverage queries set showed similar patterns, where applying RFHT performed best on all measures. The best settings of all KBs with RFHT performed better across all measures compared to the baseline with RFHT.

6 Analysis and discussion

In summary, from Table 7, we highlight the following observations:

- PRF harmed effectiveness, independent of the KB and of the PRF approach used (including the PRFHT method). While both PRF and PRFHT selected only the top ranked health terms, not all health terms in the top ranked documents were related to the query. For example, the results retrieved by query "lay down cough" (query number 104003) contained many terms related to "coughing", such as "flu". While "cough" might relate to flu, pages discussing flu may not necessarily be relevant to the original query. Hence, we found that performing PRF(HT) on expanded queries resulted in query drift, and generated results with higher residuals compared to methods without PRF(HT). Nevertheless, after residuals were reduced through the use of condensed lists (judged documents only, see Sect. 6.2.2 for the results), queries with PRF(HT) generally performed better than without PRF(HT).
- RF, instead, did provide improved effectiveness, independently of the RF approach, the KB used or the query set (high coverage of all queries).
- Both PRFHT and RFHT, which used the likelihood of expansion terms to be health related, performed generally better for all measures compared to simple PRF or RF.
- When using the all queries set and no relevance feedback, and using a combination of expansion terms from both Wikipedia and CHV (WikiCHV) performed best (on all measures). The only exception was the baseline's nDCG@10 score, which was higher. This was likely because the results obtained with WikiChv contained a higher number of unjudged documents compared to the baseline. This highlights that combining expansion terms from multiple KBs did improve the original CHS queries.
- For the high coverage queries set, expanded queries with no relevance feedback performed better than the baseline for all measures (see Table 6 (bottom)). This suggests that each KB could be used to effectively expand CHS queries. Overall, the best settings from CHV (CSE-Title) outperformed the best settings from the other KBs.
- For the high coverage queries, independently of relevance feedback, the best setting for all KBs generated a higher number of queries that produced an effectiveness gain than a loss (see Table 7 (bottom)). In fact, in these cases the gains (loss) are WSE-Title: 52.38% (38.10%), USE-Title: 47.54% (22.95%), CSE-Title: 58.33% (27.78%), and WikiChv: 54.76% (33.33%). When relevance feedback is considered (and in particular, the best feedback technique is used, i.e. RFHT), then the gain (loss) become: WSE-TitleRFHT: 68.42% (22.37%), USE-TitleRFHT: 69.74% (21.05%), CSE-TitleRFHT: 68.42% (23.68%), and WikiChvRFHT: 67.11% (23.68%).

Table 7 Influence of choices in relevance feedback (Choice 5). Statistical significance differences reported in Table 21 and 22

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert exp \rvert}$	⟨e,g,l⟩
The all queries set						
Baseline	.2465	.1798	.3263	.0399		
BaselineRF	.2055	.1777	.3412	.1400	11.70	150, 75, 74
BaselinePRF	.1657	.1704	.2679	.2831	15.63	297, 66, 146
BaselineRFHT	**.3691**	**.2336**	**.6369**	.1807	1011.90	300, 205, 65
BaselinePRFHT	.2420	.1805	.3308	.0970	25.14	300, 95, 125
GUIR-3	.1975	.1803	.2636	.2333	15.63	297, 74, 134
WSE-Title	.2430	.1843	.3231	.0824	1.37	76, 27, 32
WSE-TitleRF	.2139	.1830	.3534	.1692	10.05	183, 91, 75
WSE-TitlePRF	.1836	.1783	.2778	.2742	14.98	297, 86, 132
WSE-TitleRFHT	**.3709**	**.2335**	**.6331**	.1859	1009.99	300, 206, 64
WSE-TitlePRFHT	.2320	.1837	.3086	.1315	25.14	30, 107, 106
USE-Title	.2137	.1917	.2961	.2115	2.60	217, 67, 94
USE-TitleRF	.2429	.2206	.3705	.2031	8.93	251, 97, 85
USE-TitlePRF	.2004	.2028	.2814	.2815	22.86	300, 89, 132
USE-TitleRFHT	**.3673**	**.2323**	**.6358**	.2023	1077.53	300, 204, 68
USE-TitlePRFHT	.2119	.1904	.2769	.2666	25.14	300, 91, 124
CSE-Title	.2433	.1929	.3283	.1236	1.77	155, 59, 58
CSE-TitleRF	.2556	.2133	.3977	.1796	10.31	226, 112, 76
CSE-TitlePRF	.2223	.2004	.3218	.2041	21.86	300, 97, 123
CSE-TitleRFHT	**.3741**	**.2320**	**.6474**	.1953	1079.44	300, 202, 69
CSE-TitlePRFHT	.2403	.1931	.3231	.1782	25.14	300, 108, 110
WikiChv	.2441	.1954	.3300	.1409	1.98	181, 69, 70
WikiChvRF	.2630	.2183	.4053	.1938	9.97	237, 119, 77
WikiChvPRF	.2256	.2032	.3236	.2191	22.02	300, 102, 122
WikiChvRFHT	**.3741**	**.2328**	**.6467**	.1967	1092.15	300, 203, 69
WikiChvPRFHT	.2349	.1950	.3117	.2004	25.14	300, 107, 113
The high coverage queries set						
Baseline	.3221	.2474	.3999	.0020		
BaselineRF	.3077	.2442	.4300	.0291	12.21	43, 21, 21
BaselinePRF	.2735	.2274	.3737	.0924	15.68	76, 20, 34
BaselineRFHT	**.4775**	**.3029**	**.7183**	.0696	964.71	76, 52, 17
BaselinePRFHT	.3386	.2468	.4333	.0139	25.64	76, 33, 21
GUIR-3	.2669	.2336	.3232	.0817	15.68	76, 20, 30
WSE-Title	.3270	.2515	.4108	.0245	1.48	21, 11, 8
WSE-TitleRF	.3084	.2413	.4344	.0359	11.62	48, 25, 20
WSE-TitlePRF	.2937	.2393	.4023	.0876	14.83	76, 32, 24
WSE-TitleRFHT	**.4961**	**.3067**	**.7342**	.0616	932.00	76, 53, 16
WSE-TitlePRFHT	.3380	.2512	.4156	.0319	25.64	76, 29, 24
USE-Title	.3338	.2631	.4452	.0525	2.25	61, 29, 14
USE-TitleRF	.3642	.2954	.5005	.0602	11.63	67, 33, 13
USE-TitlePRF	.3201	.2672	.4387	.1030	21.68	76, 28, 27
USE-TitleRFHT	**.4979**	**.3036**	**.7342**	.0677	988.55	76, 52, 18

Table 7 (continued)

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert exp \rvert}$	⟨e,g,l⟩
USE-TitlePRFHT	.3298	.2587	.3964	.0574	25.64	76, 29, 26
CSE-Title	.3319	.2608	.4436	.0265	2.11	36, 21, 10
CSE-TitleRF	.3575	.2766	.4905	.0316	14.48	52, 31, 16
CSE-TitlePRF	.3178	.2590	.4200	.0546	21.30	76, 28, 28
CSE-TitleRFHT	**.4908**	**.3072**	**.7272**	.0758	980.32	76, 51, 18
CSE-TitlePRFHT	.3359	.2595	.4300	.0495	25.64	76, 30, 25
WikiChv	.3404	.2614	.4492	.0354	2.36	42, 23, 14
WikiChvRF	.3698	.2770	.5106	.0342	14.30	54, 34, 15
WikiChvPRF	.3229	.2596	.4344	.0608	21.50	76, 30, 26
WikiChvRFHT	**.4928**	**.3073**	**.7297**	.0751	991.51	76, 51, 18
WikiChvPRFHT	.3425	.2597	.4439	.0573	25.64	76, 30, 25

Bold indicates the highest effectiveness achieved for each KB

To contextualise the results obtained by the KB retrieval methods, in Table 7, we also reported the results of the method implemented by the GUIR-3 submission to the CLEF 2016 challenge (Soldaini et al. 2016). This was the best performing, comparable[9] query expansion method at CLEF 2016. The method expands queries by mapping query entities to the UMLS, then navigating the UMLS tree to gather hypernyms from mapped entities as the source of expansion. Post-processing is applied to pure entities unlikely to benefit retrieval. For each query, multiple expanded query variations are collected and their results aggregated using the Borda algorithm (see Soldaini et al. (2016) for details). Unlike the original method, our implementation relied on BM25F rather than DFR as the scoring method and QuickUMLS in place of Metamap as the entity extraction method, so as to be directly comparable with our baseline and KB retrieval methods. In Table 7, we do not report $\overline{|exp|}$ for GUIR-3 as the method replaces some of the original terms with the expansions, thus making comparisons not trivial.

While Jimmy et al. (2018) suggested that shorter expansion terms are likely to be more effective, in this study we found that is not necessarily true. Table 7 shows that the combination of Wikipedia and CHV based KB (WikiChv) has longer average expansion terms and performed better than the best settings from either Wikipedia KB or CHV based KB. Furthermore, Table 7 also shows that PRFHT and RFHT generate significantly more expansion terms and yet, they are more effective than the PRF and RF approaches.

Overall results can hide some underlying trends so we analysis the impact of query expansion on a per-query basis. Figure 5 shows the gains/losses versus baseline obtained by the best settings of Wikipedia KB (WSE-TitleRFHT), UMLS KB (USE-TitleRFHT), CHV KB (CSE-TitleRFHT), and the combination of Wikipedia and Chv KB (WikiChvRFHT). The magnitudes of these changes are shown in the figure. These improvements (or losses) were measured using RBP@10 and thus expanded queries with low coverage are unlikely to perform as effective as expanded queries with high coverage. Gains and losses were similar for the different KBs; i.e., for a given query, the gain or loss was similar irrespective of the KB. Only 5 out of the 76 high coverage queries did not exhibit this trend.

[9] ECNU-2 had the highest effectiveness, but it used Google query suggestion service to gain expansions.

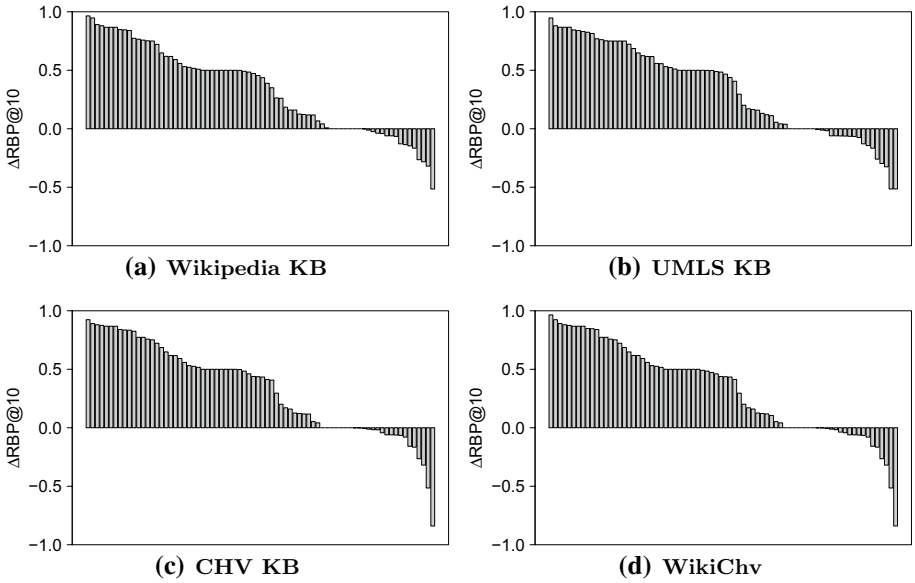

Fig. 5 Changes in RBP@10 between the Entity Query Feature Expansion model utilising the best settings versus baseline. Only high coverage queries are reported

Table 8 Performance gain/loss from expanded queries where RBP@10 gains were found in one or more KB, but losses were found in the other KBs

Query number	Wikipedia	UMLS	CHV	WikiChv
145001	0.1846	− 0.0605	− 0.0605	− 0.0605
144002	0.1601	− 0.2969	0.1601	0.1601
111004	0.0674	0.0391	− 0.8398	− 0.8398
141001	0.0078	− 0.0059	− 0.0059	− 0.0059
101006	− 0.0381	0.0557	0.0528	0.0528

Next, we investigated features of queries with expanded terms from all KBs without relevance feedback (WSE-Title, USE-Title, CSE-Title, and WikiChv). To do so, we analysed results for the high coverage queries in Choice 4 (Table 6 (bottom)) and found that of the 129 high coverage queries, 12 queries were expanded by all of the four best settings (see Table 9). The small number of overlapping expanded queries from the four best settings suggests that each best setting mostly targeted different queries. Table 9 shows similar patterns to Table 8, where gains and losses were similar for the different KBs.

Then, we investigated the 3 queries from Table 9 where mixed results were obtained across the different KBs (i.e. not all KBs consistently provided a gain (loss) for the query)—these were queries 131,002, 101,001, and 147,001. Table 10 shows that the terms added to each of the 3 queries largely differed depending on the KB used. Interestingly, Wikipedia, although being a general purpose KB, produced more relevant health expansion terms than specialised health KBs (i.e., UMLS and CHV). Nevertheless, we also found that the coverage of the Wikipedia KB was limited compared to that of the UMLS and CHV KBs. In fact, Table 6 (top) shows that the best settings that used Wikipedia KB (WSE-Title) only expanded 76 queries compared to 217 and 155 queries expanded by the

Table 9 Performance gain/loss from high coverage queries in Table 6 (bottom). Only queries that are expanded by all four best settings (WSE-Title, USE-Title, CSE-Title, and WikiChv) are reported

Query num	Wikipedia	UMLS	CHV	WikiChv
131002	0.4961	− 0.0039	− 0.0039	0.4961
147006	0.2344	0.2383	0.2383	0.4833
128004	0.2822	0.2822	0.2822	0.2822
146005	0.0147	0.4833	0.4833	0.252
131006	0.25	0.25	0.25	0.25
147005	0.2207	0.2432	0.2432	0.2432
101001	0.0224	− 0.0645	0.1894	0.1933
147004	0.1074	0.1074	0.1074	0.1074
141004	− 0.0009	− 0.0146	− 0.0009	− 0.0146
147001	0.0693	− 0.2237	− 0.2237	− 0.2237
128002	− 0.25	− 0.25	− 0.25	− 0.25
141002	− 0.0528	− 0.4424	− 0.0538	− 0.4424

Table 10 Terms added to queries 131,002: "penis lymphocytic infiltration marked nuclear crush artifact", 101,001: "inguinal hernia repair laparoscopic mesh benefits risks", and 147,001: "throat infection sore throat irritated eyes treatment options"

Query#	Wikipedia	UMLS	CHV	WikiChv
131002	Cutaneous, lymphoid, hyperplasia	Cellular	Cellular, procedure	Cutaneous, lymphoid, hyperplasia, cellular, procedure
101001	Surgery	Groin	Groin, laparoscopy, medical, subject, headings	Surgery, groin, laparoscopy, medical, subject, headings
147001	Pharyngitis	Pharyngitis, tenderness	Pharyngitis, tenderness	Pharyngitis, tenderness

best settings used for the UMLS and CHV KBs. This limitation of Wikipedia may be expected as the Wikipedia KB used in this study (WC-TypeLinks) contained only 13,135 terms—this is orders of magnitude smaller than the UMLS KB (UC-All) and CHV KB (CC-Med), which contained 3,057,234 and 1,344,941 terms, respectively.

Finally, we investigated how expansion terms from each KBs differ to each other. Table 11 shows the overlap rate among expansion terms from the best settings for all KBs. As expected, all expansion terms from Wikipedia and CHV KBs were found within the expansion terms from WikiChv. These results also further confirmed that the coverage of the Wikipedia KB was lower compared to that of the UMLS and CHV KBs. Only 3.5% of UMLS and 7.6% of CHV expansion terms were found in Wikipedia. On the other hand, 19.2% and 20.2% of expansion terms from Wikipedia were found within expansion terms from the UMLS and CHV, respectively. Finally, these results also show that each KB promoted mostly different expansion terms.

6.1 Generalisability of the best settings

We have shown that the best settings of query expansion based on Wikipedia, UMLS, CHV, or the combination of Wikipedia and CHV to form the KB, were able to improve

Table 11 The rate of overlap between expansion terms added from KB *i* with expansion terms added from KB *j*. For example, 3.5% of expansion terms from the UMLS are found in expansion terms from Wikipedia.

	Wikipedia (%)	UMLS (%)	CHV (%)	WikiChv (%)
Wikipedia	–	3.5	7.6	29.1
UMLS	19.2	–	39.9	52.0
CHV	20.2	25.4	–	76.8
WikiChv	100.0	25.35	100	–

retrieval effectiveness, compared to the original CHS queries. We did so by empirically exploring different KB retrieval settings throughout 5 choices, and selecting the best configuration for each choice. Next, we aimed to validate our findings by verifying whether they apply to a different sample of the web and a different set of CHS queries.

To this aim, we applied the best settings we obtained on the CLEF 2016 collection to the CLEF2015 collection. This collection contains 66 queries and a corpus of more than 1 million web pages, sampled from health related websites (rather than a general sample, as in CLEF 2016, i.e. Clueweb09). Table 12 reports the results obtained when applying the best settings for Wikipedia, UMLS, CHV, and the combination of Wikipedia and CHV to the CLEF2015 collection. The results showed that:

- Independently of the KB, RFHT exhibited improvement, but PRFHT did not. These findings were in line with those from CLEF2016.
- For the all queries set, without relevance feedback, expanded queries from WSE-Title, CSE-Title, and WikiChv provided gains over the baseline for bpref and RBP@10. However, other than WSE-Title, other expansion methods performed worst for nDCG@10 compared to the baseline.
- For the high coverage queries set, without relevance feedback, the best settings for CHV (CSE-Title) and for the combination of Wikipedia and CHV (WikiChv) performed better than the baseline for all measures.

In summary, the above findings show that the settings that were found to best perform on CLEF 2016 did translate to the CLEF 2015 collection.

6.2 Mitigating problems with unjudged documents

The analysis of residuals for expanded queries (top part of Tables 3, 4, 5, 6, 7), along with the analysis in Fig. 4, indicated that the baseline had far less unjudged documents amongst the top 10 results, compared with the EQFE method. We treated unjudged documents as not-relevant; however, given the shallow pools at CLEF 2016, and the fact that the method investigated here did not contribute to the pool (and is substantially different from those that did), there is the possibility that a significant portion of the unjudged documents were, in fact, relevant. To account for this in our analysis of results, along with reporting RBP residuals, we also used bpref (which only considers assessed documents) and further considered the high coverage queries sub-set for each result set (bottom part of Tables 3, 4, 5, 6, 7).

Next, we further analyse our results with respect to unjudged documents, by (1) using the additional relevance assessments made available for this collection in CLEF 2017 (Palotti et al. 2017), and (2) using condensed list evaluation measures (Sakai 2007).

Table 12 Performance of the CLEF2016's best settings for CLEF2015 queries set. Statistical significance differences reported in Table 23

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert exp \rvert}$	⟨e,g,l⟩
The all queries set						
Baseline	.2782	.2649	.3501	.0380		
BaselineRFHT	**.5559**	**.5195**	**.7789**	.1559	802.11	66, 51, 11
BaselinePRFHT	.2396	.2663	.2696	.0799	23.55	66, 13, 33
WSE-Title	.2785	.2651	.3503	.0470	1.33	3, 2, 0
WSE-TitleRFHT	**.5471**	**.5188**	**.7740**	.1578	813.55	66, 50, 12
WSE-TitlePRFHT	.2388	.2663	.2717	.0943	23.55	66, 13, 33
USE-Title	.2361	.2563	.3002	.1848	1.95	38, 7, 17
USE-TitleRFHT	**.5290**	**.4958**	**.7534**	.1802	842.77	66, 50, 12
USE-TitlePRFHT	.2047	.2561	.2462	.2724	23.55	66, 11, 35
CSE-Title	.2777	.2838	.3616	.1424	1.52	25, 6, 8
CSE-TitleRFHT	**.5396**	**.5071**	**.7689**	.1816	821.55	66, 51, 11
CSE-TitlePRFHT	.2354	.2827	.2800	.2045	23.55	66, 16, 33
WikiChv	.2769	.2827	.3618	.1420	1.50	26, 7, 8
WikiChvRFHT	**.5349**	**.5069**	**.7654**	.1828	822.42	66, 50, 12
WikiChvPRFHT	.2346	.2813	.2821	.2051	23.55	66, 16, 33
The high coverage queries set						
Baseline	.3423	.2996	.3938	.0022		
BaselineRFHT	**.6407**	**.5658**	**.8790**	.0505	765.90	20, 16, 2
BaselinePRFHT	.3254	.3008	.3246	.0111	23.80	20, 4, 11
WSE-Title	.3397	.2961	.3946	.0012	1.00	1, 1, 0
WSE-TitleRFHT	**.6255**	**.5651**	**.8673**	.0543	768.80	20, 15, 3
WSE-TitlePRFHT	.3226	.2960	.3315	.0131	23.80	20, 4, 11
USE-Title	.3058	.2791	.3636	.0152	2.30	10, 1, 5
USE-TitleRFHT	**.6225**	**.5629**	**.8554**	.0552	790.00	20, 15, 3
USE-TitlePRFHT	.2710	.2728	.2969	.0879	23.80	20, 3, 13
CSE-Title	.3488	.3118	.4084	.0150	1.29	7, 2, 3
CSE-TitleRFHT	**.6336**	**.5600**	**.8788**	.0506	781.15	20, 16, 2
CSE-TitlePRFHT	.3200	.3026	.3366	.0348	23.80	20, 5, 11
WikiChv	.3462	.3083	.4092	.0140	1.25	8, 3, 3
WikiChvRFHT	**.6184**	**.5593**	**.8671**	.0544	784.05	20, 15, 3
WikiChvPRFHT	.3172	.2978	.3435	.0369	23.80	20, 5, 11

Bold indicates the highest effectiveness achieved for each KB

Submission to CLEF 2017

We submitted results from our previous work (Jimmy et al. 2017) to the CLEF 2017 e-Health IR Task 1 (Palotti et al. 2017). In CLEF 2017, the topics from 2016, which we considered in our experiments, were re-used to obtain a deeper and more varied assessment pool. We thus further applied this new set of assessments to study the choices in knowledge based retrieval considered here. Table 13 reports the effectiveness of all expanded queries for Choice 5, using the combined relevance assessments from CLEF 2016 and 2017.

For the all queries set, the top part of Table 13 shows that queries expanded using any of the KBs studied here and without relevance feedback (i.e., WSE-Title, USE-Title, CSE-Title, or WikiChv) performed better than the baseline, on all measures, with the exception of WSE-Title (worse nDCG@10) and USE-Title (worse nDCG@10 and RBP@10).

While the evaluation results from CLEF 2017 have reduced the number of unjudged documents retrieved using expanded queries, we found that residuals from all expanded queries were consistently higher than residuals from the baseline query (see Fig. 6 for which we used the combined CLEF 2016 and 2017 relevance assessments.

We thus turn to analyse the results for the high coverage queries (Table 13, bottom part). For this set, the expanded queries based on any KB and without relevance feedback (i.e., WSE-Title, USE-Title, CSE-Title, or WikiChv) performed better than the baseline on all measures, with the exception of USE-Titlte, which had a lower nDCG@10. Overall, the results from the combined CLEF 2016 and 2017 assessments confirmed our findings as summarised at the beginning of Sect. 6.

Condensed list evaluation

Sakai (2007) suggested computing evaluation measures such as nDCG or average precision on condensed lists, i.e., document rankings obtained by considering only judged documents, as an alternative to bpref for dealing with retrieval results hampered by unjudged documents. We followed this approach for further analyse the results. In Table 14 we report the performance of queries expanded with and without relevance feedback, using condensed list evaluation for precision at 10 (P@10), mean average precision (MAP), nDCG@10 and RBP@10 (For brevity, statistical significant differences are reported in Table 26.). Condensed list results suggest that queries expanded with any KB without relevance feedback (i.e., WSE-Title, USE-Title, CSE-Title, or WikiChv) performed better than the baseline, on all measures. Any relevance feedback method (RF, PRF, RFHT, or PRFHT) could further improve retrieval effectiveness on all measures, with the only exception of applying RF and PRF to WSE-Title, which obtained a lower MAP than when used without PRF (i.e., WSE-Title > WSE-TitleRF, WSE-TitlePRF).

7 Conclusions

In this paper, we explored the influence of different choices in knowledge base (KB) retrieval for consumer health search (CHS). Choices included KB construction, entity mention extraction, entity mapping, source of expansion, and relevance feedback. We compared the effectiveness of a general KB (Wikipedia), a medical specialised KB (UMLS) and a consumer health vocabulary (CHV) as the basis for query expansion.

Our empirical evaluation (as summarised in Table 15) showed that the best settings for the Wikipedia KB are:

1. Index only Wikipedia pages that have health related infobox types or links to medical terminologies.
2. Use uni-, bi-, and tri-grams of the original queries that matched CHV terms as entity mentions.
3. Map entity mentions to Wikipedia entities based on the Aliases feature.
4. Source expansion terms from the mapped Wikipedia page Title.

Table 13 Influence of choices in KB construction for Choice 5 using the combined CLEF 2016 and 2017 relevance assessments (compare with results from Table 7, where only CLEF 2016 assessments were used). Statistical significance analysis is reported in Tables 24 and 25

Choice	nDCG@10	bpref	RBP@10	Res.	⟨e,g,l⟩
All query set					
Baseline	.2482	**.1603**	.3354	.0022	
BaselineRF	.2139	.1460	.3573	.0080	150, 75, 73
BaselinePRF	.1842	.1347	.2898	.0495	297, 90, 143
BaselineRFHT	**.3509**	.1576	**.6477**	.1461	300, 209, 63
BaselinePRFHT	.2493	.1579	.3481	.0207	300, 123, 98
WSE-Title	.2478	**.1624**	.3367	.0068	76, 31, 28
WSE-TitleRF	.2256	.1488	.3786	.0126	183, 95, 72
WSE-TitlePRF	.2008	.1439	.3065	.0700	297, 103, 133
WSE-TitleRFHT	**.3532**	.1568	**.6452**	.1446	300, 212, 60
WSE-TitlePRFHT	.2406	.1607	.3300	.0236	300, 117, 107
USE-Title	.2190	.1680	.3081	.1370	217, 72, 97
USE-TitleRF	.2489	**.1874**	.3873	.1293	251, 104, 84
USE-TitlePRF	.2096	.1738	.2955	.1870	300, 99, 131
USE-TitleRFHT	**.3511**	.1555	**.6492**	.1647	300, 210, 64
USE-TitlePRFHT	.2198	.1658	.2974	.1648	300, 104, 125
CSE-Title	.2491	.1709	.3436	.0459	155, 68, 61
CSE-TitleRF	.2667	**.1807**	.4194	.0767	226, 122, 75
CSE-TitlePRF	.2356	.1723	.3414	.1004	300, 110, 123
CSE-TitleRFHT	**.3591**	.1560	**.6619**	.1490	300, 209, 65
CSE-TitlePRFHT	.2493	.1691	.3431	.0766	300, 120, 111
WikiChv	.2516	.1714	.3460	.0448	181, 79, 70
WikiChvRF	.2739	**.1829**	.4278	.0730	237, 128, 76
WikiChvPRF	.2372	.1733	.3421	.0994	300, 111, 124
WikiChvRFHT	**.3586**	.1569	**.6621**	.1481	300, 211, 64
WikiChvPRFHT	.2449	.1702	.3341	.0799	300, 119, 114
High coverage query set					
Baseline	.2945	**.1942**	.3904	.0018	
BaselineRF	.2537	.1745	.3941	.0064	92, 40, 50
BaselinePRF	.2359	.1653	.3489	.0282	169, 52, 81
BaselineRFHT	**.4041**	.1745	**.7044**	.0761	170, 116, 39
BaselinePRFHT	.3034	.1924	.4031	.0069	170, 72, 54
WSE-Title	.2956	.1974	.3921	.0038	54, 24, 23
WSE-TitleRF	.2738	.1776	.4302	.0093	113, 56, 48
WSE-TitlePRF	.2603	.1796	.3792	.0320	169, 68, 68
WSE-TitleRFHT	**.4081**	.1746	**.7035**	.0714	170, 120, 36
WSE-TitlePRFHT	.2903	**.1978**	.3755	.0057	170, 64, 63
USE-Title	.2920	.2060	.4110	.0376	118, 51, 39
USE-TitleRF	.3264	**.2282**	.4915	.0303	142, 73, 33
USE-TitlePRF	.2801	.2179	.3932	.0781	170, 69, 64
USE-TitleRFHT	**.4097**	.1740	**.7055**	.0879	170, 114, 42
USE-TitlePRFHT	.2879	.2051	.3835	.0559	170, 68, 62
CSE-Title	.3113	.2060	.4283	.0153	75, 40, 26

Table 13 (continued)

Choice	nDCG@10	bpref	RBP@10	Res.	⟨e,g,l⟩
CSE-TitleRF	.3302	**.2190**	.4949	.0260	121, 74, 36
CSE-TitlePRF	.3005	.2109	.4138	.0358	170, 70, 66
CSE-TitleRFHT	**.4147**	.1737	**.7224**	.0788	170, 116, 40
CSE-TitlePRFHT	.3131	.2049	.4259	.0216	170, 70, 60
WikiChv	.3152	.2073	.4325	.0100	97, 51, 35
WikiChvRF	.3443	**.2225**	.5152	.0170	129, 81, 36
WikiChvPRF	.3040	.2120	.4235	.0305	170, 70, 67
WikiChvRFHT	**.4124**	.1744	**.7179**	.0815	170, 117, 40
WikiChvPRFHT	.3057	.2078	.4126	.0217	170, 69, 62

Bold indicates the highest effectiveness achieved for each KB

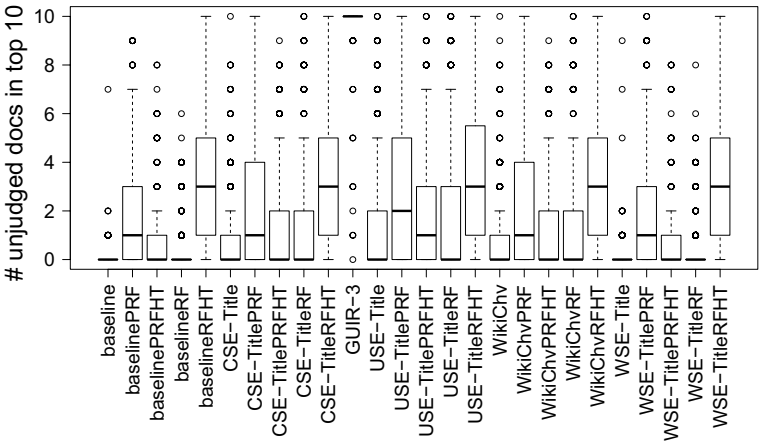

Fig. 6 Unjudged documents among the top 10 retrieved by runs in Table 13 (top)

5. Add relevance feedback terms filtered based on the likelihood of being health related (RFHT).

As for the UMLS KB, the best settings are:

1. Index all UMLS concepts.
2. Use uni-, bi-, and tri-grams of the original queries that matched UMLS terms as entity mentions.
3. Map entity mentions to UMLS entities based on the Aliases feature.
4. Source expansion terms from the mapped UMLS Title feature.
5. Add relevance feedback terms filtered based on the likelihood of being health related (RFHT).

For the CHV KB, the best settings are:

1. Index all CHV concepts that are related to the four key aspects of medical decision criteria.

Table 14 Performance of expanded queries with and without relevance feedback, using condensed list evaluation. Statistical significance differences reported in Table 26

Choice	P@10		MAP		nDCG@10		RBP@10	
Baseline	.3167		.1652		.2605		.3337	
BaselineRF	.3307	+4.4%	.1612	−2.4%	.2817	+8.1%	.4032	+20.8%
BaselinePRF	.3150	−0.5%	.1430	−13.4%	.2571	−1.3%	.3453	+3.5%
BaselineRFHT	.5210	+64.5%	.2001	+21.1%	.4864	+86.7%	.7382	+121.2%
BaselinePRFHT	.3273	+3.4%	.1650	−0.2%	.2717	+4.3%	.3579	+7.2%
WSE-Title	.3243	+2.4%	.1701	+3.0%	.2701	+3.7%	.3535	+5.9%
WSE-TitleRF	.3400	+7.4%	.1652	+0.0%	.2923	+12.2%	.4229	+26.7%
WSE-TitlePRF	.3370	+6.4%	.1528	−7.5%	.2773	+6.5%	.3686	+10.4%
WSE-TitleRFHT	**.5250**	**+65.8%**	**.2000**	**+21.0%**	**.4894**	**+87.9%**	**.7388**	**+121.4%**
WSE-TitlePRFHT	.3283	+3.7	.1679	+1.6%	.2710	+4.0%	.3569	+7.0%
USE-Title	.3227	+1.9%	.1718	+4.0%	.2699	+3.6%	.3477	+4.2%
USE-TitleRF	.3773	+19.2%	**.2058**	**+24.5%**	.3159	+21.3%	.4243	+27.2%
USE-TitlePRF	.3487	+10.1%	.1822	+10.3%	.2866	+10.0%	.3573	+7.0%
USE-TitleRFHT	**.5233**	**+65.3%**	.1994	+20.7%	**.4913**	**+88.6%**	.7519	**+125.3%**
USE-TitlePRFHT	.3263	+3.1%	.1687	+2.1%	.2738	+5.1%	.3563	+6.8%
CSE-Title	.3283	+3.7%	.1757	+6.3%	.2790	+7.1%	.3548	+6.3%
CSE-TitleRF	.3813	+20.4%	.1997	+20.8%	.3313	+27.2%	.4571	+37.0%
CSE-TitlePRF	.3592	+13.4%	.1824	+10.4%	.3028	+16.2%	.3859	+15.7%
CSE-TitleRFHT	**.5310**	**+67.7%**	**.2009**	**+21.6%**	**.4947**	**+89.9%**	**.7599**	**+127.7%**
CSE-TitlePRFHT	.3347	+5.7%	.1750	+5.9%	.2842	+9.1%	.3732	+11.8%
WikiChv	.3340	+5.5%	.1781	+7.8%	.2857	+9.7%	.3672	+10.0%
WikiChvRF	.3880	+22.5%	**.2026**	**+22.6%**	.3385	+29.9%	.4677	+40.2%
WikiChvPRF	.3632	+14.7%	.1852	+12.1%	.3055	+17.3%	.3895	+16.7%
WikiChvRFHT	**.5367**	**+69.5%**	.2011	+21.7%	**.4981**	**+91.2%**	**.7605**	**+127.9%**
WikiChvPRFHT	.3373	+6.5%	.1761	+6.6%	.2861	+9.8%	.3746	+12.3%

Bold indicates the highest effectiveness achieved for each KB

Table 15 Summary of Table 7 comparing results from the baseline and those from the best settings of each KB for all queries set

Choice	nDCG@10	bpref	RBP@10	Res.
Baseline	.2465	.1798	.3263	.0399
WSE-TitleRFHT	.3709* (+ 50.5%)	.2335* (+ 29.9%)	.6331* (+ 94.0%)	.1859
USE-TitleRFHT	.3673* (+ 49.0%)	.2323* (+ 29.2%)	.6358* (+ 94.9%)	.2023
CSE-TitleRFHT	.3741* (+ 51.8%)	.2320* (+ 29.0%)	.6474* (+ 98.4%)	.1953
WikiChvRFHT	.3741* (+ 51.8%)	.2328* (+ 29.5%)	.6467* (+ 98.2%)	.1967

The asterisks indicate statistical significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) between the baseline and the respective result

2. Use uni-, bi-, and tri-grams of the original queries that matched CHV terms as entity mentions.
3. Map entity mentions to CHV entities based on the Aliases feature.

4. Source expansion terms from the mapped CHV Title feature.
5. Add relevance feedback terms filtered based on the likelihood of being health related (RFHT).

Finally, the best combined settings are:

1. Combine expansion terms from the best settings of Wikipedia and CHV (WikiChv).
2. Add relevance feedback terms filtered based on the likelihood of being health related (RFHT).

Our empirical evaluation shows that, overall, combining expansion terms from the best settings of Wikipedia and CHV (WikiChv) was more effective than using expansion terms from the best settings of any individual KB. Using expansion terms from the combined KBs (WikiChv) improved upon the baseline in both bpref (+ 8.7%) and RBP@10 (+ 1.1%); this when using the full query set and without relevance feedback. For high coverage queries, improvements were observed for nDCG@10 (+ 5.7%), bpref (+ 5.7%), and RBP@10 (+ 12.3%). While the best results were observed using the combined, WikiChv, KB, the use of each individual KBs resulted in improvements over their respective baselines on high coverage queries. These findings demonstrate the merit of a knowledge-base retrieval approach in the challenging CHS domain.

The use of relevance feedback with filtering of health related query terms further improved results. For the full query set, expansion with a combined WikiChvRFHT KB improved considerably compared to the baseline: nDCG@10 (+ 51.8%), bpref (+ 29.5%), and RBP@10 (+ 98.2%). For high coverage queries, similar improvements were observed: nDCG@10 (+ 53%), bpref (+ 24.2%), and RBP@10 (+ 82.5%).

The major limitation of our experiments was the number of unjudged documents retrieved using the expanded queries on the CLEF 2016 collection. We addressed this limitation in different ways. When reporting the RBP results, we also reported the residuals: these provide an intuition of how much RBP could be under-estimated because of treating unjudged documents as not relevant. For each set of experiments, we considered also a subset of queries for which a larger portion of assessed documents were retrieved by all approaches. We also further augmented the set of assessed documents from CLEF 2016 with the relevance assessments for the same queries made available as part of CLEF 2017. This evaluation further confirmed the findings obtained when considering only the CLEF 2016 assessments. Finally, we also analysed the retrieval results with respect to a condensed lists-based evaluation (i.e., by considering only judged documents). The condensed list evaluation confirmed our findings that expanded queries with or without (pseudo) relevance feedback from all KB performed better than the baseline. Yet, it remains challenging to fairly evaluate the methods, because of the number of relevance assessments available in the collection. Nevertheless, this work provides an extended investigation into the choices in KB retrieval for CHS, highlighting both what worked and what did not.

Acknowledgements Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan/LPDP). Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Research Award.

Appendix 1: Statistical significance analysis

See Tables 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26.

Table 16 Statistical significance analysis for results in Table 3: Choice 1. n, b, and r mark statistical significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

	baseline	WC-All	WC-Type	WC-TypeLinks	WC-UMLS	UC-All	UC-Med	CC-All	CC-Med
baseline	-								
WC-All	nbr	-							
WC-Type	nbr	...	-						
WC-TypeLinks	nbr	...	n..	-					
WC-UMLS	nbr	...	...	...	-				
UC-All	n.r	n.r	nb.	...	.b.	-			
UC-Med	nbr	...	...	...	...	...	-		
CC-All	nbr	...	...	...	...	nb.	...	-	
CC-Med	nbr	n.r	n.r	...	n.r	...	n.r	nbr	-

The High Coverage Queries Set

	baseline	WC-All	WC-Type	WC-TypeLinks	WC-UMLS	UC-All	UC-Med	CC-All	CC-Med
baseline	-								
WC-All	n..	-							
WC-Type	.b.	.br	-						
WC-TypeLinks	...	.b.	...	-					
WC-UMLS	.b.	.b.	...	..r	-				
UC-All	.b.	nb.	n..	n..	...	-			
UC-Med	nb.	nb.	nb.	nb.	nbr	nbr	-		
CC-All	nb.	nb.	nb.	nb.	n.r	nb.	...	-	
CC-Med	nb.	nb.	nb.	nb.	nbr	nbr	.b.	.b.	-

Table 17 Statistical significance analysis for results in Table 4: Choice 2. n, b, and r mark statistical significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

	baseline	WME-All	WME-CHV	WME-UMLS	WME-Metamap	UME-All	UME-CHV	UME-UMLS	UME-Metamap	CME-All	CME-CHV	CME-UMLS	CME-Metamap
baseline	-												
WME-All	nbr	-											
WME-CHV	nbr	n.r	-										
WME-UMLS	n.r	nbr	.b.	-									
WME-Metamap	nbr	...	n.r	nbr	-								
UME-All	n.r	...	...	...	...	-							
UME-CHV	n.r	...	...	...	...	...	-						
UME-UMLS	n.r	.b.	...	...	.b.	nbr	nbr	-					
UME-Metamap	nbr	...	...	...	...	...	...	nbr	-				
CME-All	nbr	...	...	...	...	...	...	...	...	-			
CME-CHV	nbr	n..	...	...	n..	...	n..	...	n.r	n.r	-		
CME-UMLS	n.r	n..	...	...	nb.	n..	n..	...	n.r	nb.	.b.	-	
CME-Metamap	nbr	...	...	...	...	...	...	.b.	...	...	n.r	nb.	-

The High Coverage Queries Set

	baseline	WME-All	WME-CHV	WME-UMLS	WME-Metamap	UME-All	UME-CHV	UME-UMLS	UME-Metamap	CME-All	CME-CHV	CME-UMLS	CME-Metamap
baseline	-												
WME-All	...	-											
WME-CHV	...	...	-										
WME-UMLS	...	...	n..	-									
WME-Metamap	...	...	...	...	-								
UME-All	...	...	...	...	...	-							
UME-CHV	...	...	...	...	...	...	-						
UME-UMLS	...	...	...	...	...	...	...	-					
UME-Metamap	...	...	...	...	...	...	...	...	-				
CME-All	...	...	...	...	...	.b.	nb.	.b.	.b.	-			
CME-CHV	...	...	...	...	...	.b.	.b.	.b.	.b.	...	-		
CME-UMLS	...	...	...	...	...	.b.	.b.	.b.	.b.	...	...	-	
CME-Metamap	...	...	...	...	...	.b.	.b.	.b.	.b.	...	...	...	-

Table 18 Statistical significance analysis for results in Table 5 (top): Choice 3 - all queries set. n, b, and r mark statistical significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

baseline	-																					
WEM-Title	nbr	-																				
WEM-Aliases	nbr	n.r	-																			
WEM-Links	nbr	.b.	nbr	-																		
WEM-Body	nbr	...	n.r	.b.	-																	
WEM-Cat	nbr	n.r	n.r	nb.	n.r	-																
WEM-All	nbr	n..	n.r	...	...	n.r	-															
UEM-Title	n.r	.b.	n.r	.b.	.b.	.br	nb.	-														
UEM-Aliases	n.r	.br	.b.	nbr	nbr	.b.	nbr	n.r	-													
UEM-Body	nbr	n.r	nbr	n.r	n.r	n.r	nbr	nbr	nbr	-												
UEM-Parent	nbr	...	nbr	...	...	n.r	...	.b.	nbr	n.r	-											
UEM-Related	n.r	...	n.r	.b.	...	...	n..	...	nbr	nbr	.b.	-										
UEM-All	n.r	...	n.r	.b.	...	n.r	...	n..	nbr	nbr	.b.	n..	-									
UEM-QuickUmls	n.r	.b.	n.r	.b.	.b.	nbr	.b.	n..	n.r	nbr	.b.	.r	...	-								
CEM-Title	n.r	.b.	n.r	nb.	.b.	nb.	n..	..r	nbr	nb.	...	n..	n.r	-								
CEM-Aliases	n..	nbr	.b.	nbr	nbr	nbr	n.r	n.r	nbr	nbr	nbr	nbr	n.r	n.r	-							
CEM-Body	nbr	n.r	nbr	.r	...	n.r	...	nb.	nbr	n.r	...	nbr	.b.	.b.	nbr	nbr	-					
CEM-Parent	nbr	nb.	nbr	...	...	nbr	...	nb.	nbr	n.r	...	nb.	.b.	.b.	nb.	nbr	...	-				
CEM-Related	n.r	.br	n..	nbr	nbr	...	nbr	..r	...	nbr	nbr	..r	n.r	n.r	..r	n.r	nbr	nbr	-			
CEM-All	nbr	...	n.r	.b.	...	n..	...	.b.	nbr	nbr	.b.	...	...	.br	nb.	nbr	n.r	.b.	nbr	-		
CEM-QuickUmls	n.r	.b.	n.r	.b.	.b.	.br	nb.	...	n.r	nbr	.b.	...	n..	n..	n..	n.r	nb.	nb.	..r	.b.	-	
	CEM-QuickUmls	CEM-All	CEM-Related	CEM-Parent	CEM-Body	CEM-Aliases	CEM-Title	UEM-QuickUmls	UEM-All	UEM-Related	UEM-Parent	UEM-Body	UEM-Aliases	UEM-Title	WEM-All	WEM-Cat	WEM-Body	WEM-Links	WEM-Aliases	WEM-Title	baseline

Table 19 Statistical significance analysis for results in Table 5 (bottom): Choice 3 - high coverage queries set. n, b, and r mark statistically significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The High Coverage Queries Set

	baseline	WEM-Title	WEM-Aliases	WEM-Links	WEM-Body	WEM-Cat	WEM-All	UEM-Title	UEM-Aliases	UEM-Body	UEM-Parent	UEM-Related	UEM-All	UEM-QuickUmls	CEM-Title	CEM-Aliases	CEM-Body	CEM-Parent	CEM-Related	CEM-All	CEM-QuickUmls
baseline	-																				
WEM-Title	...	-																			
WEM-Aliases	...	...	-																		
WEM-Links	.b.	nb.	.b.	-																	
WEM-Body	.b.	nb.	.b.	...	-																
WEM-Cat	...	...	...	.b.	.b.	-															
WEM-All	...	...	...	n..	n..	...	-														
UEM-Title	...	...	...	...	...	n..	...	-													
UEM-Aliases	...	...	...	...	nb.	nb.	...	...	-												
UEM-Body	nbr	nbr	nbr	...	..r	nbr	nbr	nbr	nbr	-											
UEM-Parent	...	...	...	...	.b.	...	...	...	...	nbr	-										
UEM-Related	...	...	...	...	nb.	nbr	...	...	...	nbr	...	-									
UEM-All	...	...	...	...	n.r	n..	...	...	...	nbr	...	...	-								
UEM-QuickUmls	...	...	...	...	n.r	nb.	...	...	...	nbr	...	...	...	-							
CEM-Title	...	...	...	...	.b.	.b.	...	...	.b.	...	nbr	.b.	...	...	-						
CEM-Aliases	...	...	...	...	nbr	nbr	...	...	...	n.r	nbr	...	...	...	...	-					
CEM-Body	.b.	.b.	.b.	n.r	...	...	...	...	...	nbr	...	...	...	...	...	...	-				
CEM-Parent	n.r	n..	n..	...	...	n.r	n.r	n.r	n.r	.b.	n.r	n.r	n.r	n.r	n.r	nbr	n.r	n.r	-		
CEM-Related	...	...	...	...	nbr	nbr	...	...	...	nbr	...	...	...	...	...	...	...	...	n.r	-	
CEM-All	n..	nb.	...	...	..	...	n..	n..	...	br	...	nb.	nb.	nb.	nb.	nb.	n.r	...	...	nb.	-
CEM-QuickUmls	...	...	...	nbr	nb.	...	...	...	...	nbr	...	...	...	...	...	...	...	...	n.r	...	nb.

Table 20 Statistical significance analysis for results in Table 6 Choice 4. n, b, and r mark statistically significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

	baseline	WSE-Title	WSE-Aliases	WSE-All	USE-Title	USE-Aliases	USE-All	CSE-Title	CSE-Aliases	CSE-All	WikiUmlsChv	WikiUmls	WikiChv	UmlsChv
baseline	-													
WSE-Title	.b.	-												
WSE-Aliases	nbr	nbr	-											
WSE-All	nbr	nbr	...	-										
USE-Title	nbr	nb.	.b.	.b.	-									
USE-Aliases	n.r	n.r	.b.	.b.	n.r	-								
USE-All	n.r	n.r	.b.	.b.	n.r	n.r	-							
CSE-Title	.b.	.b.	nbr	nbr	n.r	n.r	n.r	-						
CSE-Aliases	n..	n..	.b.	.b.	...	n.r	n.r	nbr	-					
CSE-All	n..	n..	.b.	.b.	...	n.r	n.r	nbr	...	-				
WikiUmlsChv	nb.	.b.	nbr	nbr	n.r	n.r	nbr	n..	.b.	.b.	-			
WikiUmls	nb.	nb.	.br	.br	.b.	n.r	n.r	n.r	...	...	..r	-		
WikiChv	.b.	.b.	nbr	nbr	n.r	n.r	n.r	.b.	nbr	nbr	n..	n.r	-	
UmlsChv	nb.	nb.	nbr	nbr	..r	n.r	n.r	n..	.b.	.b.	.b.	...	n..	-

The High Coverage Queries Set

	baseline	WSE-Title	WSE-Aliases	WSE-All	USE-Title	USE-Aliases	USE-All	CSE-Title	CSE-Aliases	CSE-All	WikiUmlsChv	WikiUmls	WikiChv	UmlsChv
baseline	-													
WSE-Title	.b.	-												
WSE-Aliases	...	n..	-											
WSE-All	...	...	...	-										
USE-Title	.b.	.b.	n..	...	-									
USE-Aliases	...	...	...	...	n..	-								
USE-All	...	...	...	...	n..	...	-							
CSE-Title	nbr	nbr	n.r	n.r	...	n.r	n.r	-						
CSE-Aliases	...	...	...	...	...	n..	n..	..r	-					
CSE-All	...	...	...	...	...	n..	n..	..r	nbr	-				
WikiUmlsChv	.br	.br	n.r	n.r	...	n.r	nbr	...	...	...	-			
WikiUmls	.b.	.b.	n..	...	...	n..	n..	...	...	...	n.r	-		
WikiChv	nbr	nbr	n.r	n.r	...	n.r	n.r	...	.b.	.b.	...	...	-	
UmlsChv	.br	.br	n.r	n.r	n.r	n.r	n.r	...	...	...	...	n.r	...	-

Table 21 Statistical significance analysis for results in Table 7 (top): Choice 5 - All queries set. n, b, and r mark statistically significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

baseline	-																							baseline
baselineRF	n..	-																						baselineRF
baselinePRF	n.r	n.r	-																					baselinePRF
baselineRFHT	nbr	nbr	nbr	-																				baselineRFHT
baselinePRFHT	...	n..	n.r	nbr	-																			baselinePRFHT
GUIR-3	n.r	.r	n..	nbr	n.r	-																		GUIR-3
WSE-Title	.b.	n..	n.r	nbr	...	n.r	-																	WSE-Title
WSE-TitleRF	n..	nb.	nbr	nbr	n..	.r	n.r	-																WSE-TitleRF
WSE-TitlePRF	n.r	n.r	nb.	nbr	n.r	...	n.r	n.r	-															WSE-TitlePRF
WSE-TitleRFHT	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	nbr	-														WSE-TitleRFHT
WSE-TitlePRFHT	...	n..	n.r	nbr	nbr	n.r	...	.r	n..	nbr	-													WSE-TitlePRFHT
USE-Title	nbr	.br	nb.	nbr	nbr	.b.	nb.	.r	n..	nbr	.b.	-												USE-Title
USE-TitleRF	.br	nb.	nbr	nbr	n.r	.br	nbr	nb.	nbr	n.r	.br	nbr	-											USE-TitleRF
USE-TitlePRF	nbr	.br	nb.	nbr	nbr	.b.	nbr	.br	.b.	nbr	nb.	...	nbr	-										USE-TitlePRF
USE-TitleRFHT	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	nbr	...	nbr	nbr	n.r	nbr	-									USE-TitleRFHT
USE-TitlePRFHT	nbr	.r	nb.	nbr	nbr	...	n.r	.r	n..	nbr	nbr	...	nbr	...	nbr	-								USE-TitlePRFHT
CSE-Title	.b.	nb.	nbr	nbr	.b.	nbr	.b.	n..	n.r	nbr	.b.	n.r	.br	n.r	nbr	n.r	-							CSE-Title
CSE-TitleRF	.br	nbr	nbr	n.r	.br	nbr	nbr	nbr	n.r	nbr	nbr	.r	n.r	n.r	nbr	.br	-							CSE-TitleRF
CSE-TitlePRF	nb.	.b.	nbr	nbr	nb.	nbr	nb.	.br	nbr	nbr	.b.	...	nbr	n.r	nbr	..r	n..	nbr	-					CSE-TitlePRF
CSE-TitleRFHT	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	nbr	...	nbr	nbr	n.r	nbr	...	nbr	n.r	nbr	n.r	-				CSE-TitleRFHT
CSE-TitlePRFHT	.b.	n..	nbr	nbr	.b.	nbr	.b.	n..	n.r	nbr	.b.	n..	.br	n.r	nbr	n.r	...	.br	...	nbr	-			CSE-TitlePRFHT
WikiChv	.b.	nb.	nbr	nbr	.b.	nbr	.b.	n..	nbr	nbr	.b.	n.r	.br	n.r	nbr	n.r	.b.	.br	n..	nbr	...	-		WikiChv
WikiChvRF	.br	nbr	nbr	n.r	nbr	nbr	nbr	nbr	n.r	nbr	nbr	n.r	nbr	n.r	nbr	nbr	nb.	nbr	n.r	nbr	nbr	nbr	-	WikiChvRF
WikiChvPRF	nb.	nb.	nbr	nbr	.b.	nbr	.b.	.b.	nbr	nbr	.b.	...	.br	n.r	nbr	.r	n..	nbr	...	nbr	nbr	n..	nbr	-	WikiChvPRF
WikiChvRFHT	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	...	nbr	nbr	n.r	nbr	...	nbr	nbr	n.r	nbr	...	nbr	nbr	n.r	nbr	-	WikiChvRFHT
WikiChvPRFHT	.b.	nb.	nbr	nbr	.b.	nbr	.b.	n.r	nb.	nbr	.b.	n..	.br	n..	nbr	n.r	...	nbr	...	nbr	...	nbr	...	nbr	-	WikiChvPRFHT

Table 22 Statistical significance analysis for results in Table 7 (bottom): Choice 5 - high coverage queries set. n, b, and r mark statistically significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The High Coverage Queries Set

baseline	-																														
baselineRF	...	-																													
baselinePRF	n..	n.r	-																												
baselineRFHT	n.r	n.r	nbr	-																											
baselinePRFHT	...	...	n.r	n.r	-																										
GUIR-3	nbr	n.r	...	n.r	n.r	-																									
WSE-Title	...	...	n..	n.r	...	nbr	-																								
WSE-TitleRF	...	...	n.r	n.r	...	n.r	...	-																							
WSE-TitlePRF	...	...	.b.	n.r	n..	..r	n..	...	-																						
WSE-TitleRFHT	n.r	n.r	nbr	n.r	n.r	n.r	n.r	n.r	n.r	-																					
WSE-TitlePRFHT	...	...	n..	n.r	...	nbr	...	...	n..	n.r	-																				
USE-Title	.br	...	nbr	n.r	.b.	nbr	.b.	...	n..	n.r	...	-																			
USE-TitleRF	nbr	nbr	nbr	n.r	.br	nbr	nbr	nbr	nbr	n.r	.br	nbr	-																		
USE-TitlePRF	...	...	nbr	n.r	...	nbr	...	.b.	.b.	n.r	...	...	nbr	-																	
USE-TitleRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	n.r	n.r	...	n.r	...	n.r	n.r	-																
USE-TitlePRFHT	.b.	...	nb.	n.r	.b.	nbr	...	...	...	n.r	...	...	...	nbr	...	n.r	-														
CSE-Title	.br	...	nbr	n.r	.b.	nbr	.b.	...	n..	n.r	...	...	...	nbr	...	n.r	...	-													
CSE-TitleRF	nbr	nbr	nbr	n.r	.br	nbr	..r	nbr	nbr	n.r	..r	...	...	b.	n.r	n.r	..r	-													
CSE-TitlePRF	...	...	nbr	n.r	...	n.r	...	...	...	n.r	...	...	...	nbr	...	n.r	...	...	nbr	-											
CSE-TitleRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	n.r	n.r	...	n.r	...	n.r	n.r	n.r	...	n.r	n.r	n.r	n.r	-										
CSE-TitlePRFHT	.b.	...	nbr	n.r	.b.	nbr	...	...	n..	n.r	...	...	...	br	...	n.r	...	...	..r	...	n.r	-									
WikiChv	.br	...	nbr	n.r	.b.	nbr	.br	n..	n..	n.r	...	...	...	.br	...	n.r	...	...	...	...	n.r	...	-								
WikiChvRF	nbr	nbr	nbr	n.r	.br	nbr	n.r	nbr	nbr	n.r	..r	n.r	n.r	.b.	n.r	n.r	n.r	n.r	.r	nbr	n.r	n.r	n.r	-							
WikiChvPRF	..	...	nbr	n.r	...	n.r	...	...	...	nb.	n.r	...	...	nbr	...	n.r	...	...	...	...	n.r	...	...	...	nbr	-					
WikiChvRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	n.r	n.r	...	n.r	n.r	n.r	n.r	n.r	...	n.r	n.r	n.r	n.r	...	n.r	n.r	n.r	n.r	-				
WikiChvPRFHT	.b.	...	nbr	n.r	.b.	nbr	...	...	n..	n.r	.br	...	.br	...	n.r	.r	...	...	...	...	n.r	...	...	.r	...	n.r	-			
	baseline	baselineRF	baselinePRF	baselineRFHT	baselinePRFHT	GUIR-3	WSE-Title	WSE-TitleRF	WSE-TitlePRF	WSE-TitleRFHT	WSE-TitlePRFHT	USE-Title	USE-TitleRF	USE-TitlePRF	USE-TitleRFHT	USE-TitlePRFHT	CSE-Title	CSE-TitleRF	CSE-TitlePRF	CSE-TitleRFHT	CSE-TitlePRFHT	WikiChv	WikiChvRF	WikiChvPRF	WikiChvRFHT	WikiChvPRFHT				

Table 23 Statistical significance analysis for results for CLEF 2015 obtained using the best settings on CLEF2016 in Table 12. n, b, and r mark statistically significant differences (pairwise t-test with Bonferroni correction, $p < 0.05$) for nDCG@10, bpref, and RBP@10, respectively

The All Queries Set

	baseline	baselineRFHT	baselinePRFHT	WSE-Title	WSE-TitleRFHT	WSE-TitlePRFHT	USE-Title	USE-TitleRFHT	USE-TitlePRFHT	CSE-Title	CSE-TitleRFHT	CSE-TitlePRFHT	WikiChv	WikiChvRFHT	WikiChvPRFHT
baseline	-														
baselineRFHT	nbr	-													
baselinePRFHT	n.r	nbr	-												
WSE-Title	...	nbr	n.r	-											
WSE-TitleRFHT	nbr	...	nbr	nbr	-										
WSE-TitlePRFHT	n.r	nbr	...	n.r	nbr	-									
USE-Title	n..	nbr	...	n..	nbr	...	-								
USE-TitleRFHT	nbr	.b.	nbr	nbr	...	nbr	nbr	-							
USE-TitlePRFHT	n.r	nbr	n..	n.r	nbr	n..	..r	nbr	-						
CSE-Title	...	nbr	..r	...	nbr	..r	nbr	nbr	n.r	-					
CSE-TitleRFHT	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	-				
CSE-TitlePRFHT	...	nbr	...	n.r	nbr	...	...	nbr	nb.	n.r	nbr	-			
WikiChv	...	nbr	..r	...	nbr	..r	nbr	nbr	n.r	...	nbr	n.r	-		
WikiChvRFHT	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	-	
WikiChvPRFHT	n..	nbr	...	n..	nbr	...	...	nbr	n..	n.r	nbr	...	n.r	nbr	-

The High Coverage Queries Set

	baseline	baselineRFHT	baselinePRFHT	WSE-Title	WSE-TitleRFHT	WSE-TitlePRFHT	USE-Title	USE-TitleRFHT	USE-TitlePRFHT	CSE-Title	CSE-TitleRFHT	CSE-TitlePRFHT	WikiChv	WikiChvRFHT	WikiChvPRFHT
baseline	-														
baselineRFHT	nbr	-													
baselinePRFHT	...	nbr	-												
WSE-Title	...	nbr	...	-											
WSE-TitleRFHT	nbr	...	nbr	nbr	-										
WSE-TitlePRFHT	...	nbr	...	...	nbr	-									
USE-Title	...	nbr	...	...	nbr	...	-								
USE-TitleRFHT	nbr	...	nbr	nbr	...	nbr	nbr	-							
USE-TitlePRFHT	..r	nbr	...	..r	nbr	...	...	nbr	-						
CSE-Title	...	nbr	...	...	nbr	...	...	nbr	..r	-					
CSE-TitleRFHT	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	-				
CSE-TitlePRFHT	...	nbr	...	...	nbr	...	...	nbr	...	...	nbr	-			
WikiChv	...	nbr	...	...	nbr	...	...	nbr	..r	...	nbr	...	-		
WikiChvRFHT	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	...	nbr	nbr	-	
WikiChvPRFHT	...	nbr	...	...	nbr	...	...	nbr	...	...	nbr	...	...	nbr	-

Table 24 Statistical significance between results of the CLEF2016's best settings using CLEF2016-2017 validation data in Table 13 (top): the all queries set. n, b, and r show statistically significant (pairwise bonferroni < 0.05) for nDCG@10, bpref, and RBP@10 measure, respectively

The All Queries Set

baseline	-																						baseline
baselineRF	nb.	-																					baselineRF
baselinePRF	nbr	nbr	-																				baselinePRF
baselineRFHT	n.r	n.r	nbr	-																			baselineRFHT
baselinePRFHT	...	nb.	nbr	n.r	-																		baselinePRFHT
WSE-Title	...	nb.	nbr	n.r	...	-																	WSE-Title
WSE-TitleRF	nbr	n.r	nbr	n.r	n..	nbr	-																WSE-TitleRF
WSE-TitlePRF	nb.	.r	nb.	nbr	nbr	nbr	n.r	-															WSE-TitlePRF
WSE-TitleRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	n.r	-														WSE-TitleRFHT
WSE-TitlePRFHT	...	nb.	nbr	n.r	...	...	.br	nb.	n.r	-													WSE-TitlePRFHT
USE-Title	nbr	.br	nb.	n.r	nbr	nbr	.br	.b.	n.r	nb.	-												USE-Title
USE-TitleRF	.br	nb.	nbr	nbr	nbr	nbr	.br	nb.	nbr	nbr	nbr	-											USE-TitleRF
USE-TitlePRF	nbr	.br	nb.	nbr	nbr	nbr	.br	.b.	nbr	nb.	...	nbr	-										USE-TitlePRF
USE-TitleRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	..	n.r	n.r	nbr	nbr	-										USE-TitleRFHT
USE-TitlePRFHT	n.r	.br	nb.	n.r	nbr	n.r	.br	.b.	n.r	n.r	...	nbr	...	-									USE-TitlePRFHT
CSE-Title	.b.	nb.	nbr	n.r	.b.	.b.	nbr	nbr	n.r	.b.	n.r	.br	n.r	n.r	-								CSE-Title
CSE-TitleRF	nbr	nbr	nbr	nbr	.br	nbr	nbr	nbr	nbr	nbr	nbr	n.r	n.r	nbr	nbr	n.r	-						CSE-TitleRF
CSE-TitlePRF	.b.	nb.	nbr	n.r	.b.	...	.br	nbr	n.r	...	..r	.br	n.r	nbr	..r	...	nbr	-					CSE-TitlePRF
CSE-TitleRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	...	n.r	n.r	nbr	nbr	...	n.r	n.r	n.r	nbr	...	-				CSE-TitleRFHT
CSE-TitlePRFHT	.b.	nb.	nbr	n.r	.b.	.b.	nb.	nb.	n.r	.b.	n.r	.br	n.r	n.r	n.r	n.r	nbr	...	n.r	-			CSE-TitlePRFHT
WikiChv	.b.	nb.	nbr	n.r	.b.	.b.	nbr	nbr	n.r	.b.	n.r	.br	n.r	n.r	n.r	...	..r	n..	nbr	...	-		WikiChv
WikiChvRF	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	n.r	n.r	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	-		WikiChvRF
WikiChvPRF	.b.	nb.	nbr	n.r	.b.	...	.br	nbr	nbr	.b.	..r	.br	n.r	nbr	..r	...	nbr	...	...	nbr	-		WikiChvPRF
WikiChvRFHT	n.r	n.r	nbr	...	n.r	n.r	n.r	...	n.r	n.r	nbr	nbr	...	n.r	n.r	nbr	n.r	...	n.r	n.r	nbr	nbr	-	WikiChvRFHT
WikiChvPRFHT	.b.	nb.	nbr	n.r	.b.	.b.	nb.	n.r	.b.	n..	.br	n.r	n.r	n.r	...	n.r	...	n.r	...	nbr	...	n.r	-	WikiChvPRFHT
	baseline	baselineRF	baselinePRF	baselineRFHT	baselinePRFHT	WSE-Title	WSE-TitleRF	WSE-TitlePRF	WSE-TitleRFHT	WSE-TitlePRFHT	USE-Title	USE-TitleRF	USE-TitlePRF	USE-TitleRFHT	USE-TitlePRFHT	CSE-Title	CSE-TitleRF	CSE-TitlePRF	CSE-TitleRFHT	CSE-TitlePRFHT	WikiChv	WikiChvRF	WikiChvPRF	WikiChvRFHT	WikiChvPRFHT	

Table 25 Statistical significance between results of the CLEF2016's best settings using CLEF2016-2017 validation data in Table 13 (bottom): the high coverage queries set. n, b, and r show statistically significant (pairwise bonferroni < 0.05) for nDCG@10, bpref, and RBP@10 measure, respectively

The High Coverage Queries Set

baseline	-																						baseline
baselineRF	nb.	-																					baselineRF
baselinePRF	nbr	nbr	-																				baselinePRF
baselineRFHT	n.r	n.r	n.r	-																			baselineRFHT
baselinePRFHT	...	n..	nbr	n.r	-																		baselinePRFHT
WSE-Title	...	nb.	nbr	n.r	...	-																	WSE-Title
WSE-TitleRF	...	n.r	nbr	n.r	n..	.br	-																WSE-TitleRF
WSE-TitlePRF	n..	...	nb.	n.r	n..	n.r	-																WSE-TitlePRF
WSE-TitleRFHT	n.r	n.r	n.r	...	n.r	n.r	n.r	-															WSE-TitleRFHT
WSE-TitlePRFHT	...	nb.	nb.	n.r	.b.	...	.b.	n..	n.r	-													WSE-TitlePRFHT
USE-Title	.b.	nb.	nbr	nbr	.b.	.b.	.b.	nb.	nbr	.b.	-												USE-Title
USE-TitleRF	nbr	nbr	nbr	nbr	.b.	.b.	nb.	nbr	nbr	.b.	n..	nbr	-										USE-TitleRF
USE-TitlePRF	.b.	nb.	nbr	nbr	.b.	.b.	nb.	nb.	nbr	.b.	...	nbr	.b.	nbr	-								USE-TitlePRF
USE-TitleRFHT	n.r	n.r	n.r	..r	n.r	n.r	n.r	n.r	..r	n.r	nbr	nbr	n..	.br	n..	nbr	-						USE-TitleRFHT
USE-TitlePRFHT	.b.	nb.	nb.	nbr	.b.	.b.	nb.	nbr	nbr	.b.	n..	.br	n..	nbr	n.r	-							USE-TitlePRFHT
CSE-Title	nbr	nbr	nbr	nbr	.b.	.b.	nb.	nbr	nbr	.b.	n..	.br	n..	nbr	n.r	-							CSE-Title
CSE-TitleRF	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	.b.	n.r	nbr	nbr	nbr	-						CSE-TitleRF
CSE-TitlePRF	.b.	nb.	nbr	nbr	.b.	.b.	nb.	nb.	nbr	.b.	...	nbr	...	nbr	...	...	nbr	-					CSE-TitlePRF
CSE-TitleRFHT	n.r	n.r	n.r	..r	n.r	n.r	n.r	n.r	..r	n.r	nbr	nbr	...	nbr	n.r	nbr	nbr	nbr	-				CSE-TitleRFHT
CSE-TitlePRFHT	.b.	nb.	nbr	nbr	.b.	.b.	nb.	nbr	nbr	nbr	n..	.br	n..	nbr	n.r	...	.br	...	nbr	-			CSE-TitlePRFHT
WikiChv	nbr	nb.	nbr	nbr	.b.	nbr	nb.	nbr	nbr	nbr	n..	.br	n.r	nbr	nbr	n.r	...	.br	...	nbr	-		WikiChv
WikiChvRF	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	nbr	...	n.r	nbr	nbr	nbr	n..	nbr	nbr	nbr	nbr	-	WikiChvRF
WikiChvPRF	.b.	nb.	nbr	nbr	.b.	.b.	nb.	nbr	nbr	.b.	...	.br	n..	nbr	...	...	n.r	...	nbr	...	...	nbr	-	WikiChvPRF
WikiChvRFHT	n.r	n.r	n.r	..r	n.r	n.r	n.r	n.r	..r	n.r	nbr	nbr	nbr	...	nbr	nbr	nbr	...	nbr	nbr	nbr	nbr	nbr	-	WikiChvRFHT
WikiChvPRFHT	.b.	nb.	nbr	nbr	.b.	.b.	nb.	nbr	nbr	.br	...	.br	n..	nbr	n.r	...	nbr	...	nbr	...	...	nbr	-		WikiChvPRFHT

Table 26 Statistical significance between results using condensed evaluation in Table 14. p, m, n, and r show statistically significant (pairwise bonferroni < 0.05) for P@10, map, nDCG@10, and RBP@10 measure, respectively

baseline	-																					baseline
baselineRF	.nr	-																				baselineRF
baselinePRF	.m..	.mnr	-																			baselinePRF
baselineRFHT	pmnr	pmnr	pmnr	-																		baselineRFHT
baselinePRFHT	.n.	..r	.m..	pmnr	-																	baselinePRFHT
WSE-Title	pmnr	..r	.m..	pmnr	....	-																WSE-Title
WSE-TitleRF	p.nr	pmnr	pmnr	pmnr	..nr	.nr	-															WSE-TitleRF
WSE-TitlePRF	..r	..r	pmnr	pmnr	....	.m..	.m.r	-														WSE-TitlePRF
WSE-TitleRFHT	pmnr	pmnr	pmnr	pmnr	....	pmnr	pmnr	pmnr	pmnr	-												WSE-TitleRFHT
WSE-TitlePRFHT	....	...r	.m..	pmnr	....	....	..nr	.m..	pmnr	....	-											WSE-TitlePRFHT
USE-Title	.m..	...r	.m..	pmnr	....	....	..nr	.m..	pmnr	....	-											USE-Title
USE-TitleRF	pmnr	pmn.	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	-										USE-TitleRF
USE-TitlePRF	pmn.	.m.r	pmn.	p.nr	.m..	....	.m.r	.m..	p.nr	....	p...	pmnr	-									USE-TitlePRF
USE-TitleRFHT	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	p.nr	p.nr	-								USE-TitleRFHT
USE-TitlePRFHT	....	...r	.m..	pmnr	....	....	...r	.m..	pmnr	....	....	pmnr	....	pmnr	-							USE-TitlePRFHT
CSE-Title	.mnr	.m.r	.mn.	p.nr	.m..	....	...r	.m..	p.nr	....	....	pmnr	....	p.nr	....	-						CSE-Title
CSE-TitleRF	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	-					CSE-TitleRF
CSE-TitlePRF	pmnr	pmn.	pmnr	p.nr	.m.r	.mn.	p.nr	p.n.	p.nr	p.nr	pmnr	.m.r	...r	p.nr	p.n.	p.nr	pmnr	-				CSE-TitlePRF
CSE-TitleRFHT	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	...r	pmnr	pmnr	pmnr	pmnr	p.nr	p.nr	....	pmnr	pmnr	p.nr	p.nr	-			CSE-TitleRFHT
CSE-TitlePRFHT	pmnr	....	.mn.	pmnr	.mn.	....	...r	.m..	pmnr	.mn.	....	pmnr	....	p.nr	.m..	....	pmnr	p...	pmnr	-		CSE-TitlePRFHT
WikiChv	pmnr	.m.r	.mn.	p.nr	.m..	...r	.m..	p.nr	.mn.	..n.	pmnr	....	p.nr	.m..	.mnr	pmnr	p...	p.nr	....	-		WikiChv
WikiChvRF	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	pmnr	p.nr	p.n.	p.n.	pmnr	WikiChvRF
WikiChvPRF	pmnr	pmn.	pmnr	pmnr	m.r	pmn.	p.nr	pmn.	p.nr	.m.r	...r	p.nr	pmnr	p.nr	pmnr	pmnr	....	p.nr	p.n.	p.n.	pmnr	-	WikiChvPRF
WikiChvRFHT	pmnr	pmnr	pmnr	pmnr	pmnr	...r	pmnr	pmnr	pmnr	p.nr	....	pmnr	pmnr	pmnr	p.nr	....	pmnr	....	....	pmnr	p.nr	p.nr	WikiChvRFHT
WikiChvPRFHT	pmnr	....	.mn.	p.nr	.mn.	...r	.m..	p.nr	.mn.	..n.	pmnr	....	p.nr	.m..	....	pmnr	....	....	....	pmnr	p...	pmnr	WikiChvPRFHT

Appendix 2: List of abbreviations

	Abbreviation	Definition
General	CHS	Consumer health search
	CHV	Consumer health vocabulary
	EQFE	Entity query feature expansion
	HT	Health term
	IR	Information retrieval
	KB	Knowledge base
Methods	CC	CHV Construction
	CEM	CHV entity mapping
	CME	CHV mention extraction
	CSE	CHV source of expansion
	EM	Entity mapping
	ME	Mention extraction
	PRF	Pseudo relevance feedback
	PRFHT	Pseudo relevance feedback health term
	RF	Relevance feedback
	RFHT	Relevance feedback health term
	SE	Source of expansion
	UC	UMLS construction
	UEM	UMLS entity mapping
	UME	UMLS mention extraction
	UMLS	Unified medical language system
	USE	UMLS source of expansion
	WC	Wikipedia construction
	WEM	Wikipedia entity mapping
	WME	Wikipedia mention extraction
	WSE	Wikipedia source of expansion
Measures	<e,g,l>	<Number of expanded queries, queries with gain, queries with loss>
	$\overline{\lvert exp \rvert}$	The average number of terms added in the expanded query
	bpref	Binary preference
	MAP	Mean average precision
	nDCG@10	Normalised discounted cumulative gain at rank 10
	P@10	Precission at rank 10
	RBP@10	Rank-biased precision at rank 10
	Res.	Residual of the rank-biased precision

References

Aronson, A. R., & Lang, F. M. (2010). An overview of metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

Balaneshinkordan, S., & Kotov, A. (2016). An empirical comparison of term association and knowledge graphs for query expansion. In *European conference on information retrieval* (pp 761–767). Berlin: Springer.

Bendersky, M., Metzler, D., & Croft, W, (2012), Effective query formulation with multiple information sources. In *Proceedings of the 5th ACM international conference on web search and data mining* (pp. 443–452).

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270.

Dalton, J., Dietz, L., & Allan, J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval* (pp. 365–374).

Díaz-Galiano, M., Martín-Valdivia, M., & Ureña-López, L. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Journal of Computers in Biology and Medicine*, *39*(4), 396–403.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 8.

Fox, S., & Duggan, M. (2013). Health online 2013. Technical report. http://www.pewinternet.org/2013/01/15/health-online-2013/. Accessed 30 Oct 2018.

Jimmy, Zuccon, G., & Koopman, B. (2016). Boosting titles does not generally improve retrieval effectiveness. In *Proceedings of the 21st Australasian document computing symposium* (pp. 25–32).

Jimmy, Zuccon, G., & Koopman, B. (2017). Qut ielab at clef 2017 e-health IR task: Knowledge base retrieval for consumer health search. In *CLEF*.

Jimmy, Zuccon, G., & Koopman, B. (2018). Choices in knowledge-base retrieval for consumer health search. In *Proceedings of the 40th European conference on information retrieval*. Berlin: Springer.

Keselman, A., Smith, C. A., Divita, G., Kim, H., Browne, A. C., Leroy, G., et al. (2008). Consumer health concepts that do not map to the UMLS: Where do they fit? *Journal of the American Medical Informatics Association*, *15*(4), 496–505.

Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2006). Relating consumer knowledge of health terms and health concepts. In *Proceedings of American medical informatics association*.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., & Lawley, M. (2012). Graph-based concept weighting for medical information retrieval. In *Proceedings of the 17th Australasian document computing symposium* (pp. 80–87).

Kotov, A., & Zhai, C. (2012). Tapping into knowledge base for concept feedback: Leveraging concept net to improve search results for difficult queries. In *Proceedings of the 5th ACM international conference on web search and data mining, ACM* (pp. 403–412).

Limsopatham, N., Macdonald, C., & Ounis, I. (2013). Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th conference on open research areas in information retrieval* (pp. 1–8).

Liu, X., & Fang, H. (2015). Latent entity space: A novel retrieval approach for entity-bearing queries. *Information Retrieval Journal*, *18*(6), 473–503.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

McDaid, D., & Park, A. L. (2011). Online health: Untangling the web. Technical report. https://www.bupa.com.au/staticfiles/Bupa/HealthAndWellness/MediaFiles/PDF/LSE_Report_Online_Health.pdf. Accessed 30 Oct 2018.

Palotti, J., Goeuriot, L., Zuccon, G., & Hanbury, A. (2016). Ranking health web pages with relevance and understandability. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 965–968).

Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., & Hanbury, A. (2017). Clef 2017 task overview: The IR task at the ehealth evaluation lab. In *Working notes of conference and labs of the evaluation (CLEF) forum. CEUR workshop proceedings*.

Plovnick, R., & Zeng, Q. (2004). Reformulation of consumer health queries with professional terminology: A pilot study. *Journal of Medical Internet Research*, *6*(3), e27.

Sakai, T. (2007). Alternatives to bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07* (pp. 71–78). New York: ACM.

Silva, R., & Lopes, C. (2016). The effectiveness of query expansion when searching for health related content: Infolab at clef ehealth 2016. In *CLEF (working notes)*.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., & Frieder, O. (2015). Retrieving medical literature for clinical decision support. In *European conference on information retrieval* (pp 538–549). Berlin: Springer.

Soldaini, L., & Goharian, N. (2016). QuickUMLS: A fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR workshop, Pisa, Italy*.

Soldaini, L., & Goharian, N. (2017). Learning to rank for consumer health search: A semantic approach. In *European conference on information retrieval* (pp 640–646). Berlin: Springer.

Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., & Goharian, N. (2016). Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal*, *19*(1–2), 149–173.

Stanton, I., Ieong, S., & Mishra, N. (2014). Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 133–142).

Toms, E., & Latter, C. (2007). How consumers search for health information. *Health Informatics Journal*, *13*(3), 223–235.

Xiong, C., & Callan, J. (2015). Query expansion with freebase. In *Proceedings of the 2015 international conference on the theory of information retrieval, ACM* (pp. 111–120).

Zeng, Q., Kogan, S., Ash, N., Greenes, R., & Boxwala, A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine-Methodik der Information in der Medizin*, *41*(4), 289–298.

Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, *13*(1), 80–90.

Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, *13*(1), 24–29.

Zhang, Y. (2014). Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. *Journal of the Association for Information Science and Technology*, *65*(1), 53–68.

Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., & Butt, L. (2012). Exploiting medical hierarchies for concept-based information retrieval. In *Proceedings of the 17th Australasian document computing symposium* (pp. 111–114).

Zuccon, G., Koopman, B., & Palotti, J. (2015). Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *European conference on information retrieval MedIR'15* (pp. 562–567).

Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., & Deacon, A. (2016). The IR task at the CLEF eHealth evaluation lab 2016: User-centred health information retrieval. In *CLEF 2016-conference and labs of the evaluation forum*.