Choices in Knowledge-Base Retrieval for Consumer Health Search

Anonymised

Anonymised
author@email

Abstract. This paper investigates how retrieval using knowledge bases can be effectively translated to the consumer health search (CHS) domain. We posit that using knowledge bases for query reformulation may help to overcome some of the challenges in CHS. However, translating and implementing such approaches is nontrivial in CHS as it involves many design choices. We empirically evaluated the impact these different choices had on retrieval effectiveness. A state-of-the-art knowledge-base retrieval model — the Entity Query Feature Expansion model — was used to evaluate the following design choices: which knowledge base to use (specialised vs. generic), 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 KB 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 pitfalls and payoffs. It aims to provide some lessons to others in advancing the state-of-the-art in CHS.

1 Introduction and Related Work

A major challenge for users in consumer health search (CHS) is how to effectively represent complex and ambiguous information needs as a query [16, 13, 14]. Studies on query formulation in CHS have shown that consumers struggle to find effective query terms [14], often submitting layman and circumlocutory descriptions of symptoms instead of precise medical terms [17]. For example, people search for "skin irregularities" instead of "skin lesions" (the correct medical term for the symptom). This leads to poor retrieval effectiveness and low user satisfaction. Different approaches have been proposed to improve CHS, including query suggestion [15], learning-to-rank using syntactic, semantic or readability features [12, 7], and query expansion or reformulation [10, 9, 8].

Here we focus on overcoming the CHS problem by expanding/reformulating a health query with more effective terms (e.g., less ambiguous, synonyms, etc.). Manually replacing query terms with those from medical terminologies (e.g., UMLS) has proven effective [8]. This shows that query reformulation in the CHS can be effective — but can it be done automatically?

In the general search domain, there have been a number of automated query reformulation approaches that link queries to entities in a knowledge base (KB) such as Wikipedia and Freebase and then used these related entities for query expansion. Bendersky et al. [1] approach involved linking the query to concepts in Wikipedia. Concepts from the query, denoted κ_Q, were weighted; the same was done for concepts in each of the documents in the corpus, denoted κ_D. The relevance score $sc(Q, D)$ between query Q and document D was calculated as a relatedness measure between κ_Q and κ_D [1]. Later, the Entity Query Feature Expansion model [2] extended this by automatically expanding queries by linking them to Wikipedia. Instead of just using entities from Wikipedia (as Bendersky et al. [1] did), the Entity Query Feature Expansion model labelled words in the user query and in each document with a set of entity mentions M_Q and M_d [2]. Each entity mention was related to KB entities $e \epsilon E$, with different relationship types. The queries were expanded by including entity aliases, categories, words, and types from Wikipedia articles. The expanded query was then matched against documents in the corpus using the query likelihood model with Dirichlet smoothing.

We posit that this Entity Query Feature Expansion model would have merit in CHS. It provides a means of mapping health queries to health entities in a health related (subset of a) KB, be this either a general KB (Wikipedia) or a specialised one (e.g., UMLS). The initial query can then be expanded based on related entities. In this paper, we investigated the use of both a specialised health KB, in line with previous work that expanded queries using,e.g., MeSH or UMLS [10, 3, 9], and of a general KB like 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. Thus, we adopted the Entity Query Feature Expansion model for our empirical evaluation, determining if such a KB retrieval approach is effective for CHS.

In investigating the effectiveness of the KB retrieval approach to CHS there are a number of important design decisions. The impact of these different decisions has not been thoroughly considered when describing the proposed approach [1, 2]. Therefore, in this paper we also seek to empirically evaluate the impact of a number of different choices in KB retrieval for CHS: 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 one, or vice versa.

2 Expansion model

We implemented the Entity Query Feature Expansion model for retrieval on either both the Wikipedia and UMLS as the KB. 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

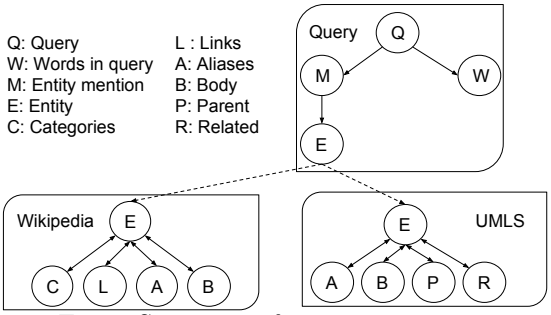


Fig. 1: Summary of expansion sources.

(B). As for the UMLS KB, a single entity is represented by the most frequently used terms for a single concept unique identifier (CUI). Features of a UMLS KB entity are aliases (A) and body (B). 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.).

3 Choices in Knowledge Base Retrieval

3.1 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 three choices for collecting health related pages: (WC-Type) based on infobox type, (WC-TypeLinks) based on infobox[1]type and links to medical terminologies as Mesh, UMLS, SNOMED CT, ICD, (WC-UMLS) based on matching between Wikipedia page titles and UMLS entities. The last method used QuickUMLS [11] to map Wikipedia page titles to the UMLS: if the mapping was successful, we included the Wikipedia entity (page) in the KB.

For UMLS KB, we considered two choices: (UC-All) all entities and (UC-Med) entities related to four key aspects of medical decision criteria (i.e., symptoms, diagnostic test, diagnoses, and treatments) as used in [6, 10]. For these choices, we included all English and non-obsolete terms.

3.2 Entity Mentions Extraction

Entity mention extraction is the process of identifying spans of text from the query that could map to some entity. It does not consider which exact entity (this

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

is detailed in the next section). We considered three possible choices to extract entity mentions: (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) [5][2]; and (ME-UMLS) include only those uni-, bi- and tri-grams of the query that matched entities in the UMLS (via QuickUMLS). These three choices were used for both the Wikipedia and UMLS KBs.

3.3 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 Figure 1, the Wikipedia entity can be represented according to six different sources; the 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 KB, the choices considered were: (UEM-Title) titles, (UEM-Aliases) aliases, (UEM-Body) the entire UMLS concept description, (UEM-Parent) parents, (UEM-Related) related entities, (UEM-All) all the previous sources (*default choice*), (UEM-QuickUmls) use QuickUMLS [11] to obtain entity mappings.

3.4 Source of Expansion

We investigated which sources in the KB were used to draw candidate terms for query expansion. We explored three choices: (SE-Title) titles of the Wikipedia pages associated with the entities, (SE-Aliases) aliases of the Wikipedia pages 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 both the Wikipedia and UMLS KBs.

3.5 Relevance Feedback

We investigated the use of relevant feedback (both explicit relevance feedback (RF) and Pseudo Relevance Feedback (PRF)). We performed RF by extracting the ten most important health related words (based on tf.idf scores) from the top three relevant documents (relevance label greater than 0). PRF was performed by extracting the ten 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 or UMLS).

4 Empirical Evaluation

To investigate the influence choices in KB retrieval have on query expansion for the CHS task, we empirically evaluated methods using the CLEF 2016

[2]Only complete string matches were considered.

eHealth [18]. 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 consider 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 [4]. 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. These candidate pages were then processed according to the choices available for KB construction (Section 3.1). Selected pages to be included in the KB were also indexed using Elasticsearch 5.1.1 with field based indexing (fields: title, links, categories, types, aliases, and body), to support the use of different fields as the source of query expansion terms (Section 3.4).

For constructing the UMLS KB, we indexed 3,057,234 non obsolete English terms 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).

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). Statistical significance ($\alpha < 0.05$) was computed using a paired t-test. 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 $< e, g, l >$.

Because of space limits, for each choice, we empirically evaluated the influence the choice had on retrieval effectiveness by examining each choice sequentially. We did this across both Wikipedia and UMLS KB, and drawed conclusions about which KB best supports CHS at the end. For each choice, we fixed the previous choices (if any) to the best setting when they were examined, and fixed the subsequent choices to their default setting. The complete set of results is provided in an online appendix at *[anonymised]*.

4.1 Knowledge base construction

The effect on retrieval of choices in KB construction is reported in Table 1 (top); results are averaged over all 300 queries in the CLEF 2016 collection. In the table, superscripts refer to statistical significance between the result and the method associated with the superscript.

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert \mathbf{exp} \rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	.2465$^{1-5}$	.1798^{1235}	.3263$^{1-5}$	.0399		
WC-Type^{1}	.0950^{0245}	.1485^{04}	.1258^{024}	.7071	38.99	299,55,161
WC-TypeLinks^{2}	**.1146**^{01}	**.1547**^{0}	**.1532**^{01}	.6361	43.22	300,66,157
WC-UMLS^{3}	.1090^{0}	.1475^{04}	.1439^{0}	.6342	21.17	299,54,163
UC-All^{4}	.1256^{01}	**.1653**^{13}	**.1626**^{01}	.5976	29.27	299,63,164
UC-Med^{5}	**.1300**^{01}	.1558^{0}	.1552^{0}	.5318	43.83	270,52,151

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert \mathbf{exp} \rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	.4481^{5}	.4700^{1345}	.5046	.0010		
WC-Type^{1}	.4567^{5}	.4160^{05}	.4342^{34}	.1736	3.54	13,5,6
WC-TypeLinks^{2}	**.4816**^{45}	**.4334**^{5}	.4944	.1129	3.54	13,5,6
WC-UMLS^{3}	.4602	.4186^{05}	**.6718**^{1}	.1814	17.54	13,9,4
UC-All^{4}	**.4285**^{25}	**.3791**^{05}	**.5874**^{1}	.0143	34.46	13,8,4
UC-Med^{5}	.3542^{0124}	.2615$^{0-4}$	.4854^{012}	.0466	46.17	12,6,5

Table 1: Influence of choices in KB construction; all queries (top) and high coverage queries (bottom).

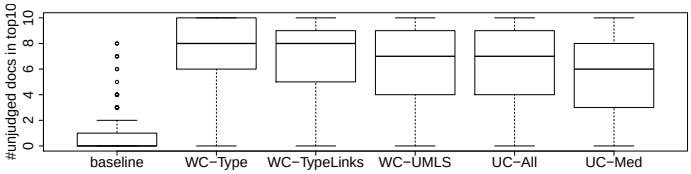


Fig. 2: Unjudged documents among the top 10 retrieved by runs in Table 1 (top).

The results for the Wikipedia KB showed that choice WC-TypeLinks (infobox type and links to medical terminologies) lead to the highest effectiveness across most measures. However, UC-All from the UMLS KB obtained higher effectiveness for all measures. Nevertheless, the baseline performed considerably better than the KB retrieval methods.

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 1. Figure 2 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 1 (compared to the residual of the baseline). Interestingly, if all unjudged documents turned out to be relevant, the RBP@10 of the KB retrieval methods would prove largely superior than that of the baseline (compare the residuals).

To further analyse the results, we considered a subset of queries for which, on average across all runs considered for a specific choice, there was a maximum of 2 unjudged documents out of the first 1. 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

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	$\langle e, g, l \rangle$
baseline^{0}	.2465$^{1-6}$	.1798^{12356}	.3263$^{1-6}$	.0399		
WME-All^{1}	**.1146**^{0}	**.1547**^{0}	.1532^{0}	.6361	43.22	300,66,157
WME-CHV^{2}	.1143^{0}	.1487^{04}	**.1573**^{0}	.6024	36.06	285,59,155
WME-UMLS^{3}	.1031^{04}	.1500^{0}	.1426^{0}	.6008	31.00	281,56,156
UME-All^{4}	**.1256**^{03}	**.1653**^{25}	**.1626**^{0}	.5976	29.27	299,63,164
UME-CHV^{5}	.1185^{0}	.1570^{04}	.1539^{0}	.6080	24.85	288,48,168
UME-UMLS^{6}	.1191^{0}	.1640^{0}	.1537^{0}	.5649	2.90	282,51,161

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\|exp\|}$	$\langle e, g, l \rangle$
baseline^{0}	.3218	.3388	.3647	.0042		
WME-All^{1}	.3795	.3286	.4516	.1874	25.32	22,9,8
WME-CHV^{2}	**.3907**^{3}	**.3295**	**.4714**	.1112	28.06	16,8,6
WME-UMLS^{3}	.3606	.3220^{2}	.4528	.0652	22.44	16,8,6
UME-All^{4}	**.3503**	.3346	**.4162**^{6}	.1488	28.62	21,12,7
UME-CHV^{5}	.3466	**.3459**	.3992	.1574	25.71	17,9,7
UME-UMLS^{6}	.3462	.3309	.3852^{4}	.1256	23.11	18,9,7

Table 2: Influence of choices in entity mention extraction; all queries (top), high coverage queries (bottom).

the considered choices (and KB); however, the subsets had the same average "coverage" with respect to the relevance assessments. We referred to these subsets as the *high coverage queries*. This subset of high coverage queries included 13 queries for choice 1 (Table 1, bottom). Results showed reduced residuals and reduced gaps between KB retrieval methods and the baselines; however trends in effectiveness across the considered choices for the Wikipedia KB did not change, unlike the relative effectiveness of the UMLS KB method (UC-All) that proved less effective than methods on the Wikipedia KB.

For Wikipedia, the results showed that the best setting was WC-TypeLinks. Thus, we selected WC-TypeLinks for the rest of the following analyses for Wikipedia KB; while we used UC-All for UMLS KB.

4.2 Entity Mentions Extraction

Table 2 (top: 300 queries and bottom: 22 high coverage queries) reports the results obtained when comparing choices for entity mention extraction. For Wikipedia, results showed that the choice of constructing entity mentions with uni-, bi- and tri-grams of the queries that matched CHV (WME-CHV) was overall the one that provided the highest retrieval effectiveness. While this is clear in the high coverage set, the difference between this strategy and using all grams (WME-All) for all queries set is less clear, probably due to the extent of many unjudged documents affecting some runs. We concluded that WME-CHV was the most effective choice and selected WME-CHV in the remaining analyses.

For UMLS, results showed that constructing entity mentions using all uni-, bi-, and tri-grams of the queries (UME-ALL) terms provided the highest retrieval effectiveness. Thus, we selected UME-ALL in the remaining analyses.

4.3 Entity Mapping

Table 3 (top: 300 queries and bottom: 22 queries) reports the results obtained when comparing choices for entity mapping. For both KBs, mapping entities

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert\mathbf{exp}\rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	.2465^{1-d}	.1798^{1-69a}	.3263^{1-d}	.0399		
WEM-Title^{1}	.1547^{02569ac}	.1602^{06789ad}	.1940^{025689}	.3699	25.60	172,32,103
WEM-Aliases^{2}	**.1984**^{0134679ab}	**.1689**^{03689a}	**.2681**$^{013-79ab}$	.2392	16.97	114,31,60
WEM-Links^{3}	.1506^{02569ac}	.1500^{0278bcd}	.2067^{0269cd}	.3130	24.23	149,22,96
WEM-Body^{4}	.1427^{025689}	.1600^{0789ad}	.1826^{02589}	.4175	71.30	204,42,121
WEM-Cat^{5}	.1783$^{013469-d}$	.1624^{06-9ad}	.2320^{0124679acd}	.2673	25.04	107,22,70
WEM-All^{6}	.1143$^{0-5789bd}$	.1487^{012578bcd}	.1573^{0123589}	.6024	36.06	285,59,155
UEM-Title^{7}	.1518^{0269ac}	.1774^{13-69a}	.1801^{02589}	.5332	16.82	287,50,160
UEM-Aliases^{8}	**.1717**$^{0469-d}$	**.1847**$^{1-69-c}$	**.2365**$^{014679-d}$	.3633	9.96	266,75,125
UEM-Body^{9}	.0734^{0-8a-d}	.1341^{0124578bcd}	.0943^{0-8a-d}	.6772	113.14	296,35,180
UEM-Parent^{a}	.1259$^{0-35789b}$	.1415^{0124578bcd}	.1702^{02589}	.5616	28.25	265,44,147
UEM-Related^{b}	.1463^{025689ac}	.1677^{3689a}	.1915^{0289d}	.5154	32.53	276,62,148
UEM-All^{c}	.1256$^{0-35789b}$	.1653^{3689ad}	.1626^{023589}	.5976	29.27	299,63,164
UEM-QuickUmls^{d}	.1355^{025689}	.1792$^{13-69ac}$	.1563^{023589b}	.5497	3.44	297,65,162

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert\mathbf{exp}\rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	.4018^{79}	.3886^{49}	.4640^{9}	.0017		
WEM-Title^{1}	.4288^{239}	**.3940**^{49}	.4715^{9}	.0559	18.86	7,4,3
WEM-Aliases^{2}	.3789^{179}	.3850^{49}	.4593^{9}	.0325	12.71	7,3,4
WEM-Links^{3}	.3655^{179b}	.3469^{9b}	.4191^{79}	.0619	33.56	9,3,6
WEM-Body^{4}	.3554^{79b}	.3289^{0125789b}	.4070^{9b}	.0328	101.77	13,4,9
WEM-Cat^{5}	.3919^{79}	.3846^{49}	.4540^{79}	.0017	3.50	2,0,2
WEM-All^{6}	**.4434**^{9}	.3711^{9}	**.5412**^{9}	.1655	24.00	15,8,6
UEM-Title^{7}	**.5051**$^{02-59}$	.3858^{49}	**.6281**^{359a}	.1612	11.10	20,11,8
UEM-Aliases^{8}	.4250^{9}	.4001^{49}	.5100^{9}	.0438	15.75	20,12,7
UEM-Body^{9}	.1752^{0-8a-d}	.2332^{0-8a-d}	.1227^{0-8a-d}	.3577	91.81	21,1,17
UEM-Parent^{a}	.3800^{9}	.3616^{9}	.4351^{79}	.2068	26.90	20,12,8
UEM-Related^{b}	.4695^{349cd}	**.4160**^{349d}	.5753^{49d}	.0564	27.10	21,14,6
UEM-All^{c}	.4114^{9b}	.3759^{9}	.5075^{9}	.1083	31.43	21,12,7
UEM-QuickUmls^{d}	.4048^{9b}	.3696^{9b}	.4615^{9b}	.1818	27.95	21,10,9

Table 3: Influence of choices in entity mapping; all queries (top), high coverage queries (bottom).

to Aliases (WEM-Aliases and UEM-Aliases) clearly outperformed the other approaches (all queries). Results for the high coverage queries showed mixed results. Thus, we selected WEM-Aliases and UEM-Aliases for the subsequent analyses.

4.4 Source of Expansion

Table 4 (top: 300 queries and bottom: 119 queries) reports the results obtained when comparing sources of query expansion. Results clearly showed that selecting titles as source of expansion (WSE-Title and USE-Title) was the most effective choice compared to other choices for both Wikipedia KB and UMLS KB. Therefore, we selected WSE-Title and USE-Title for the following analyses.

4.5 Relevance Feedback

Table 5 (top: 300 queries and bottom: 80 queries) reports the results obtained with and without relevance feedback. For Wikipedia, results showed that the

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert exp \rvert}$	$\langle e, g, l \rangle$
baseline[0]	.2465[2−6]	.1798[1−4]	.3263[2356]	.0399		
WSE-Title[1]	**.2425**[23456]	**.1843**[023]	**.3230**[2356]	.0829	1.37	76,26,32
WSE-Aliases[2]	.1976[01]	.1687[01456]	.2677[01]	.2376	16.75	114,30,61
WSE-All[3]	.1984[01]	.1689[01456]	.2681[01]	.2392	16.97	114,31,60
USE-Title[4]	**.2126**[0156]	**.1887**[023]	**.2996**[56]	.2119	2.85	235,73,98
USE-Aliases[5]	.1813[0146]	.1864[23]	.2449[014]	.3298	9.16	257,72,120
USE-All[6]	.1717[0145]	.1847[23]	.2365[014]	.3633	9.96	266,75,125

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert exp \rvert}$	$\langle e, g, l \rangle$
baseline[0]	.2794[1]	.2189[456]	.3554[1]	.0130		
WSE-Title[1]	**.2860**[0]	**.2211**[45]	**.3737**[0]	.0149	1.77	13,8,4
WSE-Aliases[2]	.2734	.2191[4]	.3645	.0446	1.82	28,17,10
WSE-All[3]	.2754	.2191[4]	.3646	.0448	11.39	28,18,9
USE-Title[4]	**.2928**[6]	**.2400**[0−3]	**.3870**	.0424	2.41	85,39,22
USE-Aliases[5]	.2633	.2357[01]	.3578	.0888	8.36	97,42,31
USE-All[6]	.2619[4]	.2346[0]	.3544	.0999	9.11	99,43,32

Table 4: Influence of choices in source of expansion; all queries (top), high coverage queries (bottom).

addition of feedback produced mixed results. RF produced the best RBP@10 across all types of queries. In terms of nDCG@10 and bpref, the Wikipedia WSE-Title choice performed better without the addition of feedback. For the UMLS, results showed that RF produced the best performance for all queries set on all measures. For the high coverage queries, the USE-Title obtained better bpref without the addition of relevance feedback. The application of relevance feedback to the baseline only improved RBP@10 when using true relevance information (RF). Nevertheless, this performed worse than the KB methods.

5 Further Analysis and Discussion

In summary, we found that: (1) PRF does not improve retrieval results, independently of the KB; (2) RF instead does provide improved effectiveness, with UMLS-based settings (USE-TitleRF) being generally better than Wikipedia-based settings (WSE-TitleRF) for both all queries and the high coverage queries sets; (3) For the high coverage queries set, independently of whether relevance feedback was applied, UMLS based KB settings were more effective than Wikipedia based KB settings; for all queries set, UMLS based KB settings with RF performed better than Wikipedia based KB settings on all measures; (4) UMLS KB expanded more queries than the Wikipedia KB. This last finding is likely due to the Wikipedia KB being incomplete in that it considered only pages with health infobox and links to medical terms. Though this was the best setting, it removed many health related pages such as "headache". Further, we found that the two methods provided radically different query expansions: on average, they only had 8.9% of expansion terms in common. On top of that, we found that they retrieved different sets of documents (average overlap for the best settings without relevance feedback: 61% (55%) of the top 1,000 (10) documents). Given these differences, we suggest future work to be directed to explore the effectiveness of combining expansions from the two KBs.

Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert\mathbf{exp}\rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	**.2465**$^{1235-9}$	.1798^{478}	.3263^{23689}	.0399		
baselineRF^{1}	.2055^{024569}	.1777^{58}	**.3412**^{23679}	.1400	11.70	150,75,74
baselinePRF^{2}	.1657^{0134578}	.1704^{578}	.2679^{01458}	.2831	15.63	297,66,146
GUIR-3^{3}	.1975^{02468}	**.1803**^{8}	.2636^{014578}	.2333	28.72	292,74,134
WSE-Title^{4}	**.2425**^{1235679}	**.1843**^{08}	.3230^{235689}	.0829	1.37	76,26,32
WSE-TitleRF^{5}	.2133^{012469}	.1833^{1268}	**.3523**^{234679}	.1710	1.02	183,92,75
WSE-TitlePRF^{6}	.1660^{0134578}	.1716^{578}	.2638^{01458}	.2928	16.17	297,71,142
USE-Title^{7}	.2126^{02469}	.1887^{0268}	.2996^{13589}	.2119	2.85	235,73,98
USE-TitleRF^{8}	**.2245**^{0236}	**.2006**$^{0-79}$	**.3687**$^{0-4679}$	.2290	9.79	263,94,93
USE-TitlePRF^{9}	.1784^{014578}	.1829^{8}	.2672^{014578}	.2989	25.35	300,70,146

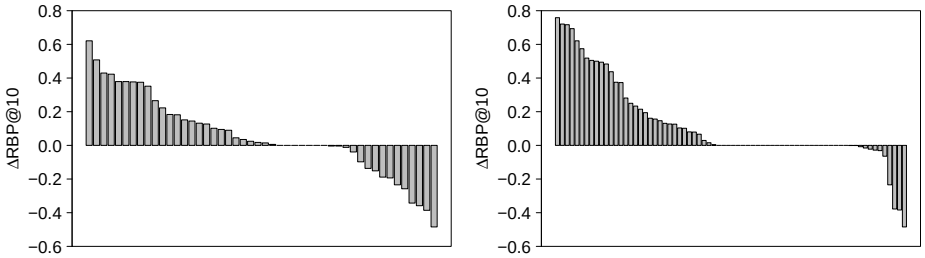
Choice	nDCG@10	bpref	RBP@10	Res.	$\overline{\lvert\mathbf{exp}\rvert}$	$\langle \mathbf{e}, \mathbf{g}, \mathbf{l} \rangle$
baseline^{0}	**.2718**^{3678}	**.2309**^{47}	.3321^{378}	.0013		
baselineRF^{1}	.2625^{68}	.2178^{78}	**.3630**^{368}	.0199	12.00	38,22,16
baselinePRF^{2}	.2429$^{6-9}$	.2142^{789}	.3339^{678}	.0662	15.66	80,25,27
GUIR-3^{3}	.2363^{04789}	.2207^{478}	.2799^{0145789}	.0875	32.87	79,22,29
WSE-Title^{4}	**.2737**^{368}	**.2378**^{037}	.3397^{378}	.0240	1.42	24,10,6
WSE-TitleRF^{5}	.2635^{68}	.2193^{78}	**.3669**^{368}	.0308	9.88	48,26,15
WSE-TitlePRF^{6}	.2272^{01245789}	.2161^{78}	.3131^{125789}	.0932	15.93	80,24,28
USE-Title^{7}	.2961^{0236}	**.2495**$^{0-6}$	.3981^{023468}	.0545	2.37	67,32,12
USE-TitleRF^{8}	**.3087**$^{0-69}$	.2445^{12356}	**.4398**$^{0-79}$	.0455	11.07	72,33,12
USE-TitlePRF^{9}	.2790^{2368}	.2323^{2}	.3748^{368}	.0800	23.19	80,30,27

Table 5: Influence of choices in relevance feedback; all queries (top), high coverage queries (bottom).

To contextualise the results obtained by KB retrieval methods, in Table 5 we also report the results of the method implemented by the GUIR-3 submission to the CLEF 2016 challenge [10]. This was the best performing, comparable[3] query expansion method at CLEF 2016. The method expands queries by mapping query entities to the UMLS, and navigating the UMLS tree to gather hypernims from mapped entities as source of expansion. Post-processing is applied to the candidates to retain expansions more likely to be of benefit to retrieval. For each query, multiple expanded query variations are collected and their results aggregated using the Borda algorithm (see [10] for details). Unlike the original method, our implementation relied on BM25F rather than DFR as scoring method and QuickUMLS in place of Metamap, so as to be directly comparable with our baseline and KB retrieval methods. In Table 5 we do not report $\overline{|exp|}$ for GUIR-3 as the method replaces some of the original terms with the expansion ones, thus making comparisons not trivial.

By observing the number of expansion terms added across the KB methods, we noted that the effective choices for KB query expansion tend to produce the lowest number of expansion terms (as well as expanding the smallest number of queries). While relevance feedback added a significant number of expansion

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



(a) Wikipedia KB (b) UMLS KB

Fig. 3: Changes in RBP@10 between the Entity Query Feature Expansion model utilising the best settings vs. baseline. Only high coverage queries are reported.

terms (as well as expanding a large number of queries), PRF did so somewhat too aggressively, which may explain why RF, which is more conservative both in queries that are expanded and the extent of expansion, outperformed PRF.

Finally, we analysed the results by considering the impact of query expansion for each query. Figure 3a reported the gains/losses vs. baseline obtained by the best settings of Wikipedia KB (WSE-TitleRF) and UMLS KB (USE-TitleRF). In total, for WSE-TitleRF (USE-TitleRF), 183 (263) queries were expanded by the model (48 (72) in the high coverage set). Of these, 15 (76) showed no change in effectiveness compared to the baseline (7 (27) in the high coverage set). Of the remaining, 92 (94) showed improvements (26 (33) in the high coverage), while 75 (93) showed losses (15 (12) in the high coverage); 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.

6 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 mentions extraction, entity mapping, source of expansion, and relevance feedback. We compared the effectiveness of a general KB (Wikipedia) and medical specialised KB (UMLS) as the basis of the query expansion. Our empirical evaluation showed that the best settings for the Wikipedia KB are: (1) indexing only Wikipedia pages that have health related infobox types or links to medical terminologies, (2) using uni-, bi-, and tri-grams of the original queries that matched CHV terms as entity mentions, (3) mapping entity mentions to the Wikipedia entity based on the Aliases feature, (4) sourcing expansion terms from the mapped Wikipedia page title, and (5) adding relevance feedback terms. As for the UMLS KB, the best settings are: (1) index all UMLS concepts, (2) using uni-, bi-, and tri-grams of the original queries that matched with UMLS title as entity mentions, (3) mapping entity mentions to the UMLS entity based on the Aliases feature, (4) sourcing expansion terms from the mapped UMLS Title feature, and (5) adding relevance feedback terms.

Results after tuning the 5 choices showed that, overall, UMLS based KB settings were more effective than Wikipedia based KB settings. The best UMLS KB settings (USE-TitleRF) performed better than the baseline in terms of bpref (+11.56%) and RBP@10 (+13%). For queries with high coverage of judged documents, USE-TitleRF was more effective for a majority of queries and outperformed the baseline on all measures: nDCG@10 (+12.58%), bpref (+5.89%), and RBP@10 (+32.43%). These results confirm that a knowledge-base retrieval approach does translate well into this often challenging CHS domain.

The major limitation of our experiments was the number of unjudged documents retrieved using the expanded queries. We mitigated this limitation by considering bpref, RBP and RBP residuals to evaluate our results. Nevertheless, this work provides the first thorough investigation of choices in KB retrieval for CHS, highlighting both pitfalls and payoffs.

References

1. M. Bendersky, D. Metzler, and W. Croft. Effective query formulation with multiple information sources. In *WSDM'12*, pages 443–452, 2012.
2. J. Dalton, L. Dietz, and J. Allan. Entity Query Feature Expansion Using Knowledge Base Links. In *SIGIR'14*, pages 365–374, 2014.
3. M. Díaz-Galiano, M. Martín-Valdivia, and L. Ureña-López. Query expansion with a medical ontology to improve a multimodal information retrieval system. *JCBM*, 39(4):396–403, 2009.
4. Jimmy, G. Zuccon, and B. Koopman. Boosting Titles Does Not Generally Improve Retrieval Effectiveness. In *ADCS'16*, pages 25–32, 2016.
5. A. Keselman, T. Tse, J. Crowell, A. Browne, L. Ngo, and Q. Zeng. Relating consumer knowledge of health terms and health concepts. In *AMIA'06*, 2006.
6. N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In *OAIR'13*, pages 1–8, 2013.
7. J. Palotti, L. Goeuriot, G. Zuccon, and A. Hanbury. Ranking health web pages with relevance and understandability. In *SIGIR'16*, pages 965–968, 2016.
8. R. Plovnick and Q. Zeng. Reformulation of consumer health queries with professional terminology: a pilot study. *JMIR*, 6(3), 2004.
9. R. Silva and C. Lopes. The effectiveness of query expansion when searching for health related content: Infolab at clef ehealth 2016. In *CLEF'16*, 2016.
10. L. Soldaini, W. Edman, and N. Goharian. Team gu-irlab at clef ehealth 2016: Task 3. In *CLEF (Working Notes)*, pages 143–146, 2016.
11. L. Soldaini and N. Goharian. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR'16*, Pisa, Italy, 2016.
12. L. Soldaini and N. Goharian. Learning to rank for consumer health search: a semantic approach. In *ECIR'17*, pages 640–646. Springer, 2017.
13. E. Toms and C. Latter. How consumers search for health information. *HIJ*, 13(3):223–235, 2007.
14. Q. Zeng, S. Kogan, N. Ash, R. Greenes, and A. Boxwala. Characteristics of consumer terminology for health information retrieval. *JMIM*, 41(4):289–298, 2002.
15. Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo, and E. Dibble. Assisting consumer health information retrieval with query recommendations. *JAMIA*, 13(1):80–90, 2006.
16. Y. Zhang. Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. *JAIST*, 65(1):53–68, 2014.
17. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *ECIR MedIR'15*, pages 562–567, 2015.
18. G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, and A. Deacon. The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In *CLEF'16*, 2016.