QUT ielab at CLEF 2017 e-Health IR Task: Knowledge Base Retrieval for Consumer Health Search

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

[1] Queensland University of Technology, Brisbane, Australia
[2] Australian E-Health Research Centre, CSIRO, Brisbane, Australia
[3] University of Surabaya (UBAYA), Surabaya, Indonesia
jimmy@hdr.qut.edu.au, g.zuccon@qut.edu.au
bevan.koopman@csiro.au

Abstract. In this paper we describe our participation to the CLEF 2017 e-Health IR Task [6]. This track aims to evaluate and advance search technologies aimed at supporting consumers to find health advice online. Our solution addressed this challenge by developing a knowledge base (KB) query expansion method. We found that the two best KB query expansion methods are mapping entity mentions to KB entities by performing exact matching entity mentions to the KB aliases (EM-Aliases) and multi-matching entity mentions to all KB features (Title, Categories, Links, Aliases, and Body) (EM-All). After mapping between entity mentions to KB entities established, we found the Title of the mapped KB entities as the best source of expansion terms compared to the aliases or combination of both features. Finally, we also found that Relevance Feedback and Pseudo Relevance Feedback are effective to further improve the query effectiveness.

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 [11, 9, 10, 13]. In this work we seek to overcome this problem by reformulating the consumer's health query with more effective terms (e.g., less ambiguous, synonyms, etc.). Previous work has shown that manually replacing query terms with those from medical terminologies (e.g., UMLS) proved to be effective [7] – but can it be done automatically?

This work addressed the adhoc search task defined in the CLEF eHealth 2017 [4], Task 3: Patient-Centred Information Retrieval, sub task IRTask1 [6]. In 2017, this task will use the same set of queries as in CLEF eHealth Task in 2016. However, only results that were un-judged in 2016 will be considered.

2 Knowledge Base Query Expansion

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 related entities for query expansion. Bendersky et al. [1] approach involved linking the query to concepts in Wikipedia. Concepts from the query, denoted as κ_Q, were weighted; the same was done for concepts in each of the documents in the corpus, denoted as κ_D. The relevance score $sc(Q, D)$ between query Q and document D was calculated as relatedness measure between κ_Q and κ_D [1].

Later, the Entity Query Feature Expansion model [2] extended this previous work by automatically expanding queries by linking them to Wikipedia. Instead of just using entities from the Wikipedia (as done by Bendersky et al. [1]), 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 general KB (Wikipedia). The initial query can then be expanded based on related entities. Our decision to use a general KB differs from other approaches in health search which typically expand the query using specialised medical KB (e.g., MeSH, UMLS) [3, 8]. Our rationale for this 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. Nevertheless, we adopt the Entity Query Feature Expansion model for our empirical evaluation, determining if such a KB retrieval approach is effective for CHS. Note however that while Wikipedia content is manually curated by an active, large community, editors may not include medical experts or clinical terminologists. Thus, there may be errors in some of the information included for medical entities in Wikipedia, also, information in Wikipedia may be incomplete.

3 Our KB Query Expansion Model for CLEF 2017

We use the Entity Query Feature Expansion model for retrieval and the Wikipedia as the KB. A single Wikipedia page represents a single entity (the page title identifies the entity). Beyond titles, Wikipedia also contains many page features useful in a retrieval scenario. Figure 1 shows those we used to map the queries to entities in the KB and as the source of expansion terms: entity title (E), categories (C), links (L), aliases (A), and body (B).

We formally define the query expansion model as:

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

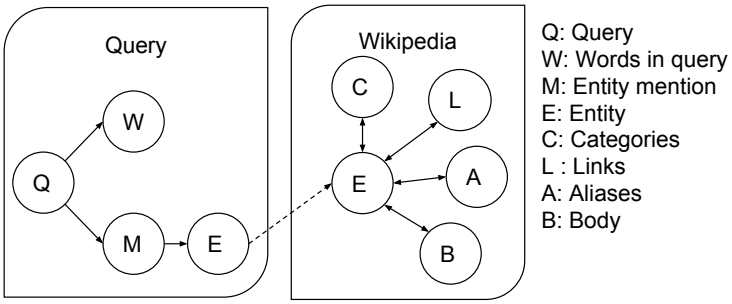

Fig. 1. Summary of expansion sources.

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 Wikipedia features EM (i.e., "Title", "Aliases", "Links", "Body", "Categories", "All") and extract expansion terms from source of expansion SE (i.e.,"Title", "Aliases", "Title and Aliases").

3.1 Relevance Feedback and Pseudo Relevance Feedback

On top of the KB query expansion, we also perform relevance feedback (RF) and Pseudo Relevance Feedback (PRF). We performed RF by extracting the ten most important health related words (based on tf.idf) from the top three relevant documents (i.e. relevance score greater than 0 in the CLEF 2016 qrel). A word is considered as health related if it exactly matches a title or an alias of a Wikipedia health page. We consider a Wikipedia page as being health related if it contains an infobox of health type or links to medical vocabulary resources, e.g MeSH.

3.2 Runs

We submitted 7 runs as described in Table 1. Runs included a baseline which consists in submitting the original, not expanded queries to a system implementing BM25F. To produce this submission, we indexed the Clueweb12b-13 collection using Elasticsearch 5.1.1, with stopping and stemming. For BM25F, we set $b = 0.75$ and $k1 = 1.2$. BM25F allows to specify 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 [5]. This is a strong baseline as it outperforms all runs submitted to CLEF 2016 (excluding the organisers' relevance feedback baselines) [12].

For constructing the KB, we considered candidate pages from the English subset of Wikipedia (dump 1/12/2016), limited to current revisions only and

Run Id	Description
1	Baseline with Relevance Feedback
2	EM-Aliases
3	EM-Aliases with Relevance Feedback
4	EM-Aliases with Pseudo Relevance Feedback
5	EM-All
6	EM-All with Relevance Feedback
7	EM-All with Pseudo Relevance Feedback

Table 1. Runs description

Run	nDCG@10	bpref	RBP@10	RBP res.
1. baselineRf	0.2117^{24567}	0.1994^{25}	$\mathbf{0.3477}^{24567}$	0.1450
2. EM-Aliases	$\mathbf{0.2357}^{134567}$	0.1835^{1346}	0.3175^{134567}	0.1060
3. EM-AliasesRf	0.2135^{24567}	$\mathbf{0.2021}^{25}$	0.3397^{4567}	0.1816
4. EM-AliasesPrf	0.1799^{1237}	0.2015^{257}	0.2680^{12357}	0.3172
5. EM-All	0.1720^{123}	0.1835^{1346}	0.2269^{12346}	0.4014
6. EM-AllRf	0.1822^{1237}	0.1954^{25}	0.2771^{12357}	0.3878
7. EM-AllPrf	0.1597^{12346}	0.1887^{4}	0.2264^{12346}	0.4668

Table 2. Performance of the runs submitted to CLEF 2017 - evaluated using CLEF 2016 relevance assessments. Superscripts refer to statistical significance between the result and the method associated with the superscript.

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 filtered by retaining only pages that contain health infobox type and links to medical terminologies as Mesh, UMLS, SNOMED CT, ICD. This choice is proven to be more effective that retaining all Wikipedia pages. The retained pages were then 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.

Once the Knowledge Base was constructed, we extended the initial query by firstly extract mentions all uni-, bi-, and tri-grams of the queries). Next, we mapped the extracted mentions to KB's entities by exact matching the query mentions to terms in KB's aliases field (EM-Aliases) and to all KB's fields (EM-All). Finally, we extended the initial query with the title of the mapped entities.

We further extended the queries from EM-Aliases and EM-All by performing Relevance Feedback (RF) and Pseudo Relevance Feedback (PRF). Our RF used the top ten health related words from the top three relevant results. Health related words are words that match title of a Wikipedia health pages (i.e., title of a page in KB). Relevant results are documents that are judged relevant following CLEF2016 qrels. In this work, PRF used the top ten health words from the top three results (regardless of whether it was judged or not).

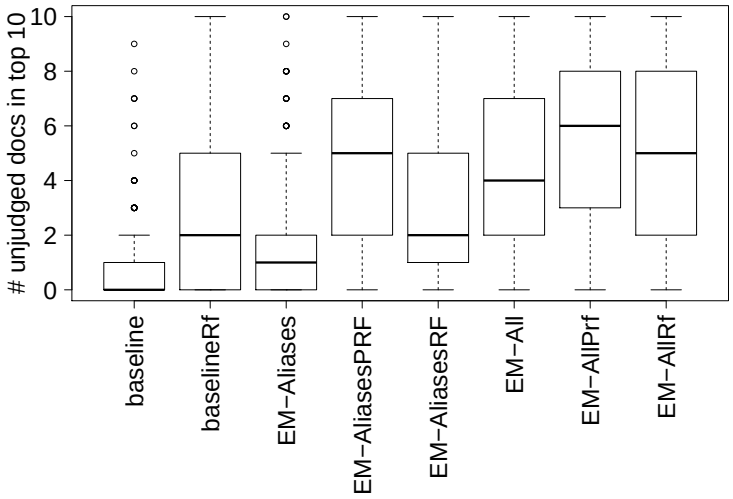

Fig. 2. Distribution of unjudged documents in top ten search results.

4 Results

Runs produced with the methods outlined above were stripped of any documents assessed in CLEF 2016, as per instructions for the CLEF 2017 submissions [6]. Before the removal of these documents, we did evaluate the results with respect to NDCG@10, BPref and RBP@10. Note that BPref results are based on the top 1,500 results for each query (this is because of the need to retrieved more documents than the 1,000 documents threshold so that when removing documents assessed in CLEF 2016, we still could retain 1,000 documents). Results according to the CLEF 2016 relevance assessments are reported in Table 2.

We further analysed the runs with respect to the number of un-judged documents retrieved (using the CLEF 2016 relevance assessments). Figure 2 shows that our expansion retrieved many un-judged documents in the top 10 search results. This observation, along with the large RBP residuals reported in Table 2, suggest that the evaluation of our runs may be affected by the large number of un-judged documents. The new assessments in CLEF 2017 may provide a fairer estimate of the effectiveness of the considered KB query expansion approaches.

5 Future Work and Conclusion

Future work will seek to further improve the effectiveness of the expanded queries by exploring post-processing the results, for example by promoting documents that are more likely to be health related.

In conclusion, using CLEF 2016 dataset, we found that Entity Query Feature Expansion Model [2] can effectively improved the query effectiveness. The expanded queries can then be further improved by performing Relevance Feedback and Pseudo Relevance Feedback.

Acknowledgment: Jimmy conducted this research as part of his doctoral study which is sponsored by Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP).

References

1. Bendersky, M., Metzler, D., Croft, W.: Effective query formulation with multiple information sources. In: WSDM'12. pp. 443–452 (2012)
2. Dalton, J., Dietz, L., Allan, J.: Entity Query Feature Expansion Using Knowledge Base Links. In: SIGIR'14. pp. 365–374 (2014)
3. Díaz-Galiano, M., Martín-Valdivia, M., Ureña-López, L.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Journal of Computers in Biology and Medicine 39(4), 396–403 (2009)
4. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab Overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (2017)
5. Jimmy, Zuccon, G., Koopman, B.: Boosting Titles Does Not Generally Improve Retrieval Effectiveness. In: ADCS'16. pp. 25–32 (2016)
6. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: 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 (2017)
7. Plovnick, R., Zeng, Q.: Reformulation of consumer health queries with professional terminology: a pilot study. JMIR 6(3) (2004)
8. Silva, R., Lopes, C.: The effectiveness of query expansion when searching for health related content: Infolab at clef ehealth 2016. In: CLEF'16 (2016)
9. Toms, E., Latter, C.: How consumers search for health information. Health Informatics Journal 13(3), 223–235 (2007)
10. Zeng, Q., Kogan, S., Ash, N., Greenes, R., Boxwala, A.: Characteristics of consumer terminology for health information retrieval. Journal of Methods of Information in Medicine 41(4), 289–298 (2002)
11. Zhang, Y.: Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. JAIST 65(1), 53–68 (2014)
12. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In: CLEF'16 (2016)
13. Zuccon, G., Koopman, B., Palotti, J.: Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In: Advances in Information Retrieval, pp. 562–567. Springer (2015)