QUT IElab at CLEF 2018 Consumer Health Search 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 2018 Consumer Health Search Task, sub task IRTask1. This track aims to evaluate and advance search technologies aimed at supporting consumers to find health advice online. Our solution addressed this challenge by extending the Entity Query Feature Expansion model (EQFE), a knowledge base (KB) query expansion method. In previous work we showed that Wikipedia, UMLS and CHV can be effective as basis for CHS query expansions within the EQFE model. To obtain the query expansion terms, first, we mapped entity mentions to KB entities by performing exact matching. After mapping, we used the Title of the mapped KB entities as the source for expansion terms. For our first three expanded query sets, we expanded the original queries sourcing expansion terms from each of Wikipedia, the UMLS, and the CHV. For our fourth expanded query set, we combined expansion terms from Wikipedia and CHV.

1 Introduction

The CLEF 2018 Consumer Health Search (CHS) Task aims to retrieve information relevant to people seeking health advice on the web [11, 13], and is a continuation of the similar task in CLEF 2017 [3, 8], but with a new, more focused document corpus in place of the more general Clueweb12B document corpus. To address this task we applied and extended the Entity Query Feature Expansion model (EQFE), a knowledge base (KB) query expansion method [2], which we have recently found performing competitively on the previous CLEF e-Health IR challenges [5]. By producing query expansions using EQFE, we seek to overcome the common issue of poor query formulation in CHS; EQFE does so by reformulating the consumer's health query with more effective terms (e.g., less ambiguous, synonyms, etc.).

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 health related terms is the Unified Medical Language

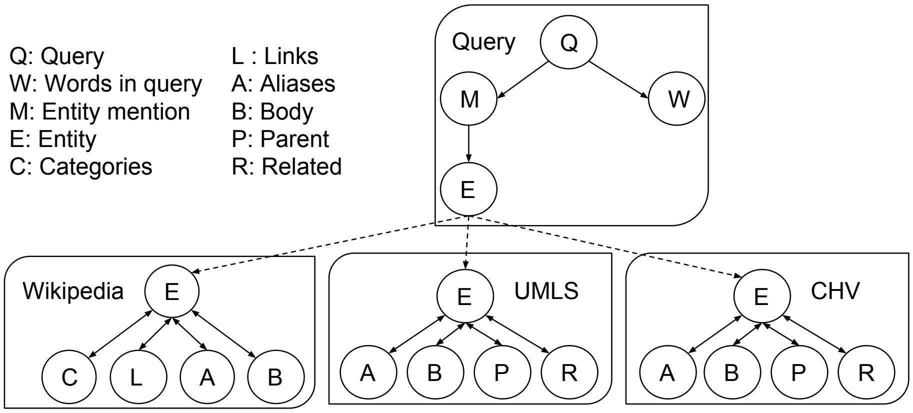

Fig. 1. Summary of expansion sources for our extension of the EQFE model.

System (UMLS) [1] – in our approach we use the UMLS as one of the sources for query expansion. However, UMLS concepts are rarely mentioned in consumer health queries: Keselman et al. [6] showed that only 8.1% of the possible n-grams constructed from consumer queries can be mapped (i.e., exact match) to UMLS concepts.

In constrast, Wikipedia is a crowdsourced, general purpose KB allowing people to promote and describe new concepts or augment existing concepts. While general purpose, Wikipedia contains considerable and detailed health information that has been effectively used in health related information retrieval [5, 9] – in our approach we use Wikipedia as one of the sources for query expansion.

In addition to UMLS and Wikipedia, we also used the Consumer Health Vocabulary (CHV) [12, 6] which was built to provide a mapping between consumer health terms and UMLS concepts. This mapping was 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 [6]. From 2007, the CHV is available as part of the UMLS entries with CHV” as source (i.e., tuples in table MRCONSO with attribute “SAB” equal to “CHV”).

2 Our KB Query Expansion Model for CLEF 2018

We implemented the Entity Query Feature Expansion model for retrieval on the Wikipedia, UMLS, and CHV 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 (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

Run Id	Source of Expansion Terms
1	The title of Wikipedia KB entities
2	The title of UMLS KB entities
3	The title of CHV KB entities
4	The combination of expansion terms from Wikipedia and CHV KBs

Table 1. Summary of the runs submitted to CLEF 2018 CHS, IRTask1.

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.).

Description of Runs

We submitted 4 runs as described in Table 1. To produce this submission, we indexed the CLEF2018 corpus using Elasticsearch 5.1.1, with stopping and Porter stemming. As underlying retrieval model we used BM25F, with $b_{title} = 0.90$, $b_{body} = 0.45$ and $k1 = 1.2$ as these settings were found to be optimal for the CLEF 2016 eHealth collection. Further, 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 the CLEF 2016 eHealth collection [4] – and we hope these values do translate well into the new CLEF 2018 CHS collection.

To obtain **Run 1**, we:

1. indexed Wikipedia pages with Medicine infobox type and pages with infobox containing links to medical terminologies such as Mesh, UMLS, SNOMED CT, etc.
2. extracted uni-, bi-, tri-grams of the original query that matched CHV entities.
3. exact matched the extracted n-grams to the Wikipedia’s aliases.
4. used the title of the matched entities as expansion terms

To obtain **Run 2**, we:

1. indexed all English and non-obsolete UMLS concepts.

	Run 1	**Run 2**	**Run 3**	**Run 4**
Run 1	-	0.0424	0.2503	0.2886
Run 2	0.0424	-	0.1706	0.1610
Run 3	0.2503	0.1706	-	0.9046
Run 4	0.2886	0.1610	0.9046	-

Table 2. Pairwise Kendall's τ_b rank correlation coefficient between runs.

2. extracted uni-, bi-, tri-grams of the original query that matched entities in the UMLS (via QuickUMLS [10]).
3. exact matched the extracted n-grams to the UMLS's aliases.
4. used the title of the matched entities as expansion terms.

To obtain **Run 3**, we:

1. indexed English and non-obsolete CHV concepts that associated to the four key aspects of medical decision criteria (i.e., symptoms, diagnostic test, diagnoses, and treatments) as used in [7].
2. extracted uni-, bi-, tri-grams of the original query that matched entities in the CHV.
3. exact matched the extracted n-grams to the CHV's aliases.
4. used the title of the matched entities as expansion terms.

To obtain **Run 4**, we combined expansion terms obtained from the Wikipedia and CHV KBs (run 1 and run 3). Using CLEF 2016 collection, we found that this combination performed the best when compared to other possible combinations.

3 Discussion

Table 2 shows the pairwise Kendall's τ_b correlation coefficient between our runs. To compute the correlation coefficient score between two runs, first, we combined query-document pairs from both runs and retain only the unique pairs. Then, we determined the rank of each document pair in each run. Query-document pairs with the same score (i.e., ties), were assigned their minimal rank. If a query-document pair was not found in one run, then the query-document pair will be assigned rank 1001 (i.e. the full length of the ranking, plus one). Finally, we used R to compute the pairwise Kendall's τ_b correlation coefficient between rank list from both runs [1].

As shown in Table 2, rank correlations among our runs are generally low (no correlation) – with the only exception of Run 3 and Run 4 which are instead

[1] https://github.com/jimmyoentung/RunsCorrelation

highly (positively) correlated. This may have been because queries in Run 4 may have been expanded using mostly by terms from the CHV KB (as used in Run 3).

4 Conclusions

In this working notes paper we have discussed the methods used by the QUT IElab team in their participation to the CLEF 2018 Consumer Health Search task (subtask 1 – ad-hoc retrieval). We submitted a total of four runs; evaluation results are not available at this stage.

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

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32(suppl_1), D267–D270 (2004)
2. Dalton, J., Dietz, L., Allan, J.: Entity Query Feature Expansion Using Knowledge Base Links. In: SIGIR'14. pp. 365–374 (2014)
3. 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: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 291–303. Springer (2017)
4. Jimmy, Zuccon, G., Koopman, B.: Boosting Titles Does Not Generally Improve Retrieval Effectiveness. In: ADCS'16. pp. 25–32 (2016)
5. Jimmy, Zuccon, G., Koopman, B., et al.: Choices in knowledge-base retrieval for consumer health search. In: European Conference on Information Retrieval. pp. 72–85. Springer (2018)
6. Keselman, A., Smith, C.A., Divita, G., Kim, H., Browne, A.C., Leroy, G., Zeng-Treitler, Q.: 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 (2008)
7. Limsopatham, N., Macdonald, C., Ounis, I.: Inferring conceptual relationships to improve medical records search. In: Proceedings of the Tenth Conference on Open Research Areas in Information Retrieval. pp. 1–8 (2013)
8. Palotti, J., Zuccon, G., Jimmy, P.P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 Task Overview: The IR task at the ehealth evaluation lab. In: CEUR-WS Proceedings (2017)
9. Soldaini, L., Cohan, A., Yates, A., Goharian, N., Frieder, O.: Retrieving medical literature for clinical decision support. In: European Conference on Information Retrieval. pp. 538–549. Springer (2015)
10. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, sigir (2016)

11. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J., Jimmy, Zuccon, G.: Overview of the clef ehealth evaluation lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer (September 2018)
12. Zeng, Q.T., Tse, T.: Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association 13(1), 24–29 (2006)
13. Jimmy, Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the CLEF 2018 Consumer Health Search Task. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2018)