Team GU-IRLAB at CLEF eHealth 2016: Task 3

Luca Soldaini, Will Edman, and Nazli Goharian

Georgetown University
Washington, DC, USA
luca@ir.cs.georgetown.edu, wce5@georgetown.edu,
nazli@ir.cs.georgetown.edu

Abstract. Recent surveys have shown that a growing number internet users seek medical help online. Yet, recent research [12] has shown that many commercial search engine still struggle in completely satisfying the information need of users. In this work, we present a study on the use of medical terms for query reformulation. We use synonyms and hypernyms from a large medical ontology to generate alternative formulations for a query; Results obtained by the reformulated queries are fused using the Borda rank aggregation algorithm.

Keywords: medical information retrieval, query reformulation, Borda rank aggregation

1 Introduction

As reported by a 2013 Pew Survey [4], growing numbers of internet users look for medical advice on the Internet, many with little or no medical experience. However, search systems fail to bridge between the layman terms of Internet users describing their conditions (e.g., *"lump with blood spots on nose"*) and the illness or disorder they are afflicted by (e.g., *"basal cell carcinoma"*), as shown by Zuccon, et al. [12].

In this manuscript, we present out efforts at the 2016 CLEF eHealth Information Retrieval Task [13]. We proposed a system that generates alternative formulations of each query using the Unified Medical Language System[1] (UMLS). UMLS has been previously exploited to process medical content generated by laypeople in information retrieval (e.g., [2]), question answering (e.g., [8]), and information extraction (e.g., [11]) tasks. Thus, we take advantage of it in our system.

In detail, for each query, we use synonyms and hypernyms extracted from UMLS to produce alternative formulations (example in Table 1); then, we retrieve results for each generated query; finally, we combine the retrieved results using Borda rank aggregation algorithm [3]. This approach was chosen due to its encouraging performances on the 2015 CLEF eHealth Task 2 dataset [9] and on a small set of query results annotated by the authors.

[1] https://www.nlm.nih.gov/research/umls/

original query	**reformulation using UMLS hypernyms**	**reformulation using UMLS synonyms**
infant labored breathing and tight wheezing cough	infant labored breathing and tight wheezing *pulmonary / upper respiratory disease*	infant labored *respiration* and tight wheezing cough

Table 1. Example of query reformulation using synonyms and hypernyms from UMLS. The expressions in *italics* have been used to replace the underlined concepts in the original query.

2 Methodology

In this section, we detail our methodology. A summary of the runs submitted to the shared task is shown in Table 2.2.

2.1 Query Reformulation

As previously mentioned, we reformulate each query using synonyms and antonyms from the UMLS metathesaurus. To identify concepts in the query, we use MetaMap [1], a tool extracting medical concepts from text documents and mapping them specific UMLS concepts. To prevent query drift in our modified queries, we considered UMLS concepts from 16 semantic types that are typically associated with the *four aspect of the medical decision criteria* (namely symptoms, diagnostic tests, diagnoses, and treatments) as suggested by Limsopatham, et al. [7].

For each expression e_i that is linked to a concept c_i in UMLS, we consider the set of atoms associated with c_i as candidate synonyms for e_i. To obtain hypernyms for an expression e_i associated with concept c_i, we use UMLS relationships database to obtain any concept c_j such that there exists a relationship of type PAR[2] between c_i and c_j to the concept.

It is often the case that synonym and hypernym identified through UMLS are quite similar to each other. This is due to the design of UMLS, which favors redundancy over correctness in aggregating multiple thesauri. To prevent duplicate queries, we use the Porter stemmer to ensure that no two added terms have the same stem, and we only add terms with an edit distance greater than four from other added terms.

To further prevent query drift, we reformulate the query only using those synonyms and hypernyms that have been deemed useful. Usefulness was estimated by considering the inverse document frequency (*idf*) of each expression in the collection [5]. Only those expressions whose *idf* is greater than 4 are used to modify the original query. Finally, we limit the number of modified queries for each concept to 8 and omit substitute expression with *idf*>11, as extremely rare synonym/hypernym concepts are less likely to find relevant results.

[2] A PAR edge signifies that the returned concepts is a parent, or hypernym, to the original concept.

IRTask	Run	Preprocessing	Query Reformulation	Rank Aggregation
IRTask1	GUIR_EN_RUN1*	stemming + case folding	*n/a*	*n/a*
IRTask1	GUIR_EN_RUN2	stemming + case folding	UMLS synonyms	Borda
IRTask1	GUIR_EN_RUN3	stemming + case folding	UMLS hypernyms	Borda
IRTask2	GUIR_EN_RUN1*	stemming + case folding	*n/a*	Borda
IRTask2	GUIR_EN_RUN3	stemming + case folding	UMLS synonyms	Borda
IRTask2	GUIR_EN_RUN3	stemming + case folding	UMLS hypernyms	Borda

Table 2. Summary of the runs submitted for evaluation. Runs identified by * represent the required baselines.

Finally, once all the reformulated queries are generated, we submit each one of them to a search engine and retrieve up to 1000 results for each one. We used the Terrier index kindly provided by the organizers to retrieve relevant documents. We decided to use Poisson model with Laplace after-effect and normalization 2 (PL2) Divergence from Randomness (DFR) model for scoring the queries, as it has been shown to be very effective for tasks that require early precision [6, 10].

2.2 Rank Aggregation

We use the Borda rank aggregation algorithm to combine results retrieved by modified queries. In detail, for each modified query, each retrieved document is given a score that is equal to the number of documents ranked below it. The total score for each document is computed by summing the score for the document for each modified query, and the aggregate ranking is created by placing each score in descending order. For task 2, we use Borda ranking to combine results from all queries in a topic.

While we experimented with other forms of rank aggregation on the 2015 CLEF eHealth (average rank, Kemeny rank aggregation), we ultimately decided to use Borda for all three submitted runs as it yields the best precision and nDCG after ten retrieved documents.

3 Experimental Results

As the ground truth for this task was not available at submission time, we could not provide any experimental results for the proposed approach.

4 Acknowledgments

This work was partially supported by the National Science Foundation through grant CNS-1204347 and REU award IIP-1362046.

References

1. A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
2. A. B. Can and N. Baykal. MedicoPort: A medical search engine for all. *Computer methods and programs in biomedicine*, 86(1):73–86, 2007.
3. C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.
4. S. Fox and M. Duggan. Health Online 2013. `http://www.pewinternet.org/Reports/2013/Health-online.aspx`, 2013.
5. D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics.* Springer, 2012.
6. B. He and I. Ounis. Term frequency normalisation tuning for bm25 and dfr models. In *Advances in Information Retrieval*, pages 200–214. Springer, 2005.
7. N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, 2013.
8. L. Nie, M. Akbari, T. Li, and T.-S. Chua. A joint local-global approach for medical terminology assignment. 2014.
9. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. CLEF, 2015.
10. V. Plachouras and I. Ounis. Usefulness of hyperlink structure for web information retrieval. In *Proceedings of ACM SIGIR*, 2004.
11. A. Yates and N. Goharian. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *Advances in Information Retrieval*, pages 816–819. Springer, 2013.
12. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can. In *Advances in Information Retrieval (ECIR)*, pages 562–567. Springer, 2015.
13. 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 2015 user-centred health information retrieval. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes.* CLEF, CEUR-WS, September 2016.