Answering PICO Clinical Questions: a Semantic Graph-Based Approach

Eya Znaidi, Lynda Tamine, Chiraz Latiri

Answering PICO Clinical Questions: a Semantic Graph-Based Approach

Eya Znaidi[1], Lynda Tamine[1], and Chiraz Latiri[2]

[1] IRIT, University Paul Sabatier of Toulouse
[2] Computer Sciences Department, Faculty of Sciences of Tunis

Abstract. In this paper, we tackle the issue related to the retrieval of the best evidence that fits with a PICO (Population, Intervention, Comparison and Outcome) question. We propose a new document ranking algorithm that relies on semantic based query expansion bounded by the local search context to better discard irrelevant documents. Experiments using a standard dataset including 423 PICO questions and more than $1,2$ million of documents, show that our aproach is promising.

Keywords: Evidence-based-Medicine, PICO Clinical Queries, Medical information retrieval, Semantic Query Expansion

1 Introduction

Evidence-Based Medicine (EBM) has been defined as the conscientious, explicit and appropriate use of systematic research findings, in consultation with the patient, with the aim of optimizing the healthcare decision-making process of medical professionals [8]. One major issue faced by the professionals during the daily practice of EBM is the complexity of expressing precise, context-specific clinical queries that better facilitate the identification of the relevant evidence. These works rely heavily on a prior automatic annotation of PICO facets in both queries and documents. Unlikely, our approach (1) relaxes the condition of PICO facet identification in the documents, and (2) abstracts the word-based question formulation by highlighting the overall semantic picture of each question facet. Moreover, each question facet is separately expanded using concepts extracted from top ranked documents issued from the initial retrieval.
The remainder of the paper is structured as follows. In section 2, we first give an overview of related work. Section 3 details our approah for PICO question elicitation and answering. In section 4, we describe the experimental setup and then present and discuss the results obtained using a standard clinical information retrieval dataset. Section 5 concludes the paper.

2 Related Work

While some previous work [1, 4, 11] tackled the issue of PICO element detection, as a prior stage before retrieving relevant documents, other studies, close to our work [3, 2, 5] focused on the design of retrieval techniques and models that exploit the PICO facets in order to compute the relevance score of documents. To achieve this goal, Boudin et al. [3, 2] automatically detected PICO elements in the documents and then revised the basic version of the IR language model [9]. More precisely, the authors revised the word-document weighting shema by taking into account both the distribution of PICO elements in the different passages of the documents and the distributions of the words in the different PICO parts. The experimental evaluation held on a collection of 1.5 million of documents and 423 queries showed that the proposed model yields an improvement of 28% in mean average precision over state-of-the-art baselines. Demner-Fushman and Lin [5] also proposed an unified framework for both detecting and using the detected PICO elements in the relevance document scoring function S_{EBM}. The latter is based on the linear combination of partial relevance scores of the documents considering the three facets of EBM, namely, PICO (S_{PICO}), strenght of evidence (S_{SoE}) and task type (S_{task}). For instance, the PICO score relies on a linear combination of P, I, C and O facet scores considering the word overlap between the document and the question. Experiments carried out on 24 real-world clinical questions show that the approach outperforms a traditional PubMed search.

3 A Semantic Graph-based Approach for Answering PICO Questions

We describe in Figure 2 the general algorithm (main and function) for expanding the PICO query and ranking the best evidence to be returned as an answer to the clinician.

- *(Main) Steps 1-12*: Given a word-based PICO query Q, the related annotation Q_{PICO}, the subqueries Q_P, Q_{IC} and Q_O and the list of N_d top ranked documents D_N^* included in a document collection C, the algorithm builds first the semantic sub-graphs G_P, G_{IC} and G_O after (1) extracting, using our concept method extraction [6] build upon Metamap[3], the active concepts, respectively $Concepts(Q_P)$, $Concepts(Q_{IC})$ and $Concepts(Q_O)$; each active concept has an importance score $Score(c)$ that highlights the likelihood of similarity between the concept preferred entry and the query words, (2) building the associated graphs G_P, G_{IC} and G_O by appending to the active concepts the corresponding hypernyms through terminology function $HypG$ processed on medical terminology T until reaching the first common concept. Each returned active concept c is considered at relative level 0.

- *(Main) Steps 13-15*: for each sub-graph G_P, G_{IC} and G_O, we build the set of N_c concepts to be used for query expansion by applying function $Expand(G_x)$ considering $Maxlevel$ which denotes the maximum level used for query expansion beginning from level 0.
- *(Expand) Steps 1-17:* To build the set of candidate concepts $Cexpand$, we consider each document d in D_N^* and then (1) extract the set of common

[3] http://metamap.nlm.nih.gov

1: Main: *Document ranking*

Input: $Q, Q_{PICO}, T, N_d, N_c, MaxLevel$
Output: G_P, G_{IC}, G_O, D_N^*
1: # Initial search
2: $D_N^* \leftarrow Top_D(Q, N_d, C)$;
3: # Query Graph Building
4: $Q_P \leftarrow Substr(Q, P)$;
5: $Q_{IC} \leftarrow Substr(Q, IC)$;
6: $Q_O \leftarrow Substr(Q, O)$;
7: $Concepts(Q_P) \leftarrow Extract(Q_P, T)$;
8: $G_P \leftarrow HypG(Concepts(Q_P), T)$;
9: $Concepts(Q_{IC}) \leftarrow Extract(Q_{IC}, T)$;
10: $G_{IC} \leftarrow HypG(Concepts(Q_{IC}), T)$;
11: $Concepts(Q_O) \leftarrow Extract(Q_O, T)$;
12: $G_O \leftarrow HypG(Concepts(Q_O), T)$;
13: $Q_P^e \leftarrow Expand(G_P)$;
14: $Q_{IC}^e \leftarrow Expand(G_{IC})$;
15: $Q_O^e \leftarrow Expand(G_O)$;
16: $Words(Q^e) \leftarrow Words(Q) \cup Entries(Q_P^e) \cup Entries(Q_{IC}^e) \cup Entries(Q_O^e)$;
17: # Final search
18: $D_N^* \leftarrow Top_D(Q^e, N_d, C)$;

2: Function: *Expand*

Input: G_x
Output: $Cexpand$
1: # Query expansion
2: # Process the top ranked documents
3: **for all** $d \in D_N^*$ **do**
4: # Extraction of document concepts
5: $Cexpand \leftarrow Extract(d, G_x)$;
6: $level \leftarrow 0$;
7: # Score Propagation
8: **for all** $c \in Cexpand$ AND $level < Maxlevel$ **do**
9: **for all** $csub \in Hypo(c, G_x)$ **do**
10: $Score(csub) \leftarrow (Score(csub) + Lev(csub) * Score(c))$;
11: $Score(csub) \leftarrow Normalized(Score(csub))$;
12: $level \leftarrow level + 1$;
13: **end for**
14: **end for**
15: **end for**
16: $Cexpand \leftarrow Top_C(G_x, N_c)$;
17: **return** $Cexpand$;

Fig. 1: The document retrieval process

weighted concepts with G_x (where $x \in \{P, IC, O\}$) using the same concept extraction method [6]; (2) apply a score propagation algorithm that propagates the scores of the active concepts of each query sub-graph G_x from level 0 to level $Maxlevel$ by iteratively summing the scores of the hyponym concepts through sub-graph G_x, $Hypo(c, G_x)$. The basic underlying idea is to leverage the importance and the specificity of the concepts by assigning the normalized scores $Normalized(Score(c))$ obtained step by step from less specific concepts to most specific ones, considering their level $Lev(c)$. The final score of a concept reflects its importance in the whole top ranked documents in terms of high specificity and matching degree with documents D_N^*. This fits with our intuition that favors the selection of most specific concepts that better match the search context gathered from the top ranked documents.

– *(Main) Steps 16-18*: The returned set of N_c top weighted concepts $Cexpand$ extracted from each sub-graph G_x, are used to expand respectively the sub-queries Q_P, Q_{IC} and Q_O (resulting in Q_P^e, Q_{IC}^e and Q_O^e respectively) by adding to the intial word-based query Q the words belonging to their preferred entries ($Entries(Q_P^e)$, $Entries(Q_{IC}^e)$ and $Entries(Q_O^e)$ respectively) within terminology T. The final expanded query Q^e is processed and allows selecting the final list of documents D_N^* to be returned as an answer to the initial PICO query Q.

4 Experimental Evaluation

4.1 Experimental Setup

We used the CLIREC dataset which has been built with the specific aim of evaluating clinical information retrieval [3]. Some statistical characteristics of the collection are depicted in Table 1.We used the MeSH terminology which has been widely accepted as the main controlled vocabulary used to index biomedical citations [10]. Each node of the terminology represents a concept node referred to using a preferred entry.

Number of documents	1.212.040 abstracts from PubMed
Average document length	246 words
Number of queries	423
Average number of query keywords	4.3 words
Average PICO query length	18.7 words
Average Number of relevant documents per query	19

Table 1: CLIREC test collection statistics

For the purpose of evaluating and comparing retrieval effectiveness, we used under version 4.0 of the Terrier search engine[4]: 1) The Mean Average Precision (MAP) measure which is the mean of the AP measure over a set of queries; it is used to provide a single, overall measure of search performance. The performance measures have been computed using the standard TREC-eval tool[5]; 2) We used two state-of-the-art information retrieval models, namely the Okapi probabilistic model (BM25) [7] and the language model (LM) [9]. The Okapi model was parameterized as recommended in the literature: $k1 = 1.2$, $k3 = 7$ and $b = 0.75$. For the LM, the Dirichlet smoothing method with $\mu = 1000$ was used.

4.2 Results

We compared the retrieval effectiveness based on MAP of our semantic graph-based document ranking algorithm *GQE* with respect to the state-of-the-art ranking models *BM25* and *LM*. Table 2 presents the obtained results in terms of the MAP measure and relevant retrieved documents as well as the corresponding pourcents of improvement and significance t values of the statistical t-test. We can see that our model (GQE) significantly overpasses word-based document ranking approaches ($BM25, LM$) from $25,44\%$ to $27,94\%$. From these results, we can highlight that our semantic approach allows achieving better results than state-of-the-art word-based IR models that do not specifically take into account the PICO framework; this yiels a credit to our intuition behind question elicitation on the basis of the semantic hidden behind each question facet.

[4] http://www.terrier.org

[5] http//trec.nist.gov/trec_eval

Model	MAP	%Change	t	Rel. Ret	% Change
$BM25$	0.1073	+25.44%	**	4783	+15.28%
LM	0.1052	+27.94%	**	4685	+17.69%
GQE	0.1346	-	-	5514	-

Table 2: Comparison of the semantic graph-based query expansion impact on the retrieval effectiveness. $\%Chg$: Student test significance over the MAP measure *: $0.01 < t \leq 0.05$; **: $0.001 < t \leq 0.01$; ***: $t \leq 0.001$.

5 Conclusion

In this paper, we presented a novel approach to answer PICO clinical queries. The key underlying idea is to enhance each query facet with the most representative terminological concepts on the basis of a local search context. Moreover, we apply a score propagation algorithm that allows selecting the concepts with higher matching degree over the whole search context and across the different query facets. Experiments on a standard data set highlight that the proposed approach significantly overpasses state-of-the-art IR models. In future, we plan to integrate a weighting facet schema in the document ranking model in order to consider the differences in the importance of the question facets with respect to document relevance.

References

1. F. Boudin, J.-Y. Nie, J. Bartlett, R. Grad, P. Pluye, and M. Dawes. Combining classifiers for robust pico element detection. *BMC Medical Informatics and Decision Making*, 10(1):29, 2010.
2. F. Boudin, J. Y. Nie, and M. Dawes. Clinical information retrieval using document and PICO structure. In *NAACL HLT*, pages 822–830, 2010.
3. F. Boudin, J.-Y. Nie, and M. Dawes. Positional language models for clinical information retrieval. In *EMNLP*, pages 108–115, 2010.
4. G. Y. Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC Med Inform Decis Mak*, 9:10, 2009.
5. D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Comput. Linguist.*, 33(1):63–103, 2007.
6. D. Dinh and L. Tamine. Towards a context sensitive approach to searching information based on domain specific knowledge sources. *Web Semantics: Science, Services and Agents on the World Wide Web*, 12(0), 2012.
7. S. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society on Informtion Science and Technology*, 27(3), 1976.
8. D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.
9. F. Song and W. B. Croft. A general language model for information retrieval. In *ACM SIGIR*, pages 279–280, 1999.
10. N. Stokes, Y. Cavedon, and J. Zobel. Exploring criteria for succesful query expansion in the genomic domain. *Information retrieval*, 12:17–50, 2009.
11. J. Zhao, M. yen Kan, P. M. Procter, S. Zubaidah, W. K. Yip, and G. M. Li. Improving search for evidence-based practice using information extraction. *BMC Medical Informatics and Decision Making*, 10(29), 2010.