Key Concept Identification for Medical Information Retrieval

Jiaping Zheng
College of Information
and Computer Sciences
University of Massachusetts
Amherst, MA
jzheng@cs.umass.edu

Hong Yu
Bedford VA Medical Center
Bedford, MA
Department of Quantitative Health Sciences
University of Massachusetts
Worcester, MA
hong.yu@umassmed.edu

Abstract

The difficult language in Electronic Health Records (EHRs) presents a challenge to patients' understanding of their own conditions. One approach to lowering the barrier is to provide tailored patient education based on their own EHR notes. We are developing a system to retrieve EHR note-tailored online consumer oriented health education materials. We explored topic model and key concept identification methods to construct queries from the EHR notes. Our experiments show that queries using identified key concepts with pseudo-relevance feedback significantly outperform (over 10-fold improvement) the baseline system of using the full text note.

1 Introduction

Allowing patients access to their own electronic health records (EHRs) can enhance medical understanding and provide clinical relevant benefits (Wiljer et al., 2006), including increased medication adherence (Delbanco et al., 2012). However, EHR notes present unique challenges to the average patients. Since these notes are not usually targeted at the patients (Mossanen et al., 2014), languages that may be difficult for non-medical professionals to comprehend are prevalent, including medical terms, abbreviations, and medical domain-specific language patterns. The valuable and authoritative information contained in the EHR is thus less accessible to the patients, who ultimately stand to benefit the most from the information.

To address the challenges, we are developing a system to link medical notes to targeted education materials from trusted resources. The textual narratives in the EHR notes are not conducive to effective and efficient query. We therefore explored topic model and key concept identification methods to construct short queries from the EHR notes.

2 Related Work

Patent retrieval (Fujii et al., 2007) is similar to this work, as the queries are usually long and complex patent documents. Several methods have been proposed to construct shorter queries from the documents. For example, words in the summary section of a patent document can be ranked by TFIDF scores and extracted to form a query (Xue and Croft, 2009). Sentences that are similar to pseudo-relevant documents according to a language model are also used to reduce query length (Ganguly et al., 2011). Other similarity measures such as Kullback-Leibler divergence are used to extract terms, which are expanded to generate queries in the patent retrieval domain (Mahdabi et al., 2012). However, the patent retrieval domain is recall-driven, while in our scenario, patients are generally not expected to read relevant education documents exhaustively.

Various methods have also been proposed to retrieve documents relevant to passages of text or web documents. A model extended from CRF is proposed to identify noun phrases and named entities from a user-selected passage as queries (Lee and Croft, 2012). Similarly, noun phrases in a verbose query are also used as candidates for key concepts (Bendersky and Croft, 2008). Other related work that reduces long queries includes ranking all subsets of the original query (Kumaran and Carvalho, 2009). However, typical EHR notes are longer than the passages and verbose queries in these systems, which makes the graphical model and other learning based models less efficient. Moreover, parsers as required by Bendersky et al. (2009) and Named Entity Recognizers for the medical domain are less effective than the general domain.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 579–584,
Lisbon, Portugal, 17-21 September 2015.

TREC Clinical Decision Support Track[1] is an IR challenge to link medical cases to information relevant for patient care. Unlike our system, this task is designed to address the physicians' information needs of diagnosis, testing and treatment for the patients.

3 Materials and Methods

3.1 Data

MedlinePlus[2] provides current and reliable information about over 900 health topics pages and 1000 medication pages to users in consumer-oriented lay language. Additionally, the medical encyclopedia section includes over 7000 articles about diseases, tests, symptoms, injuries, and surgeries. We include in this study the textual narratives in the "health topics", "drugs, supplements, and herbal information", and "medical encyclopedia" sections of the MedlinePlus as the collection of educational materials. There are a total of approximately 9400 articles in this collection, which we designate as *MedlinePlus*. Table 3.1 summarizes the characteristics of the collection.

We index the *MedlinePlus* documents with Galago, an advanced open source search engine. Galago implements the inference network retrieval model (Turtle and Croft, 1991). This model calculates the probability of the user's information needs being satisfied given a document in a directed acyclic graph. This framework is applied in many information retrieval tasks, and shown to be successful (Metzler et al., 2004).

Twenty progress notes are randomly selected from a corpus of de-identified EHR notes as the EHR document collection. Each note contains on average 261 tokens, with a standard deviation of 133. A physician read each note, and manually identified relevant education materials from the *MedlinePlus* documents. Each EHR note is linked to 22 education material documents on average. For example, Table 3.1 shows the summary of one EHR note and some of its relevant *MedlinePlus* documents. There are approximately 30 sentences or 360 tokens in the actual document.

We adopted Mean Average Precision (MAP) and Precision at 10 to evaluate the IR systems.

[1] http://www.trec-cds.org/

[2] http://www.nlm.nih.gov/medlineplus/

Summary of EHR Note
Patient remains in ICU with the following problems: respiratory failure, hemodynamics, renal failure, status post liver transplant, atrial fib, infectious disease, nutrition.
Select Relevant Documents
Respiratory Failure
Deep Vein Thrombosis
Aspiration pneumonia
Pulmonary Hypertension
Kidney Failure
Atrial Fibrillation or Flutter
Liver Transplantation
Dialysis - Hemodialysis

Table 2: Example EHR Note and its relevant documents

3.2 IR Systems

We investigate several query generation approaches, whereby short queries are built from an EHR note. In our queries, sequential dependence model (Metzler and Croft, 2005) was used to capture the dependencies in a multi-word query term. In this model, given a query, documents are ranked based on features of documents containing a single query term, two query terms sequentially appearing in the query, and two query terms in any order. This model has been shown to be effective in many applications (Balasubramanian et al., 2007; Cartright et al., 2011; Bendersky et al., 2009).

3.2.1 Baseline Approach

Our baseline approach issues each of the 20 test EHR notes as the query to the *MedlinePlus* document index and retrieves top 500 relevant documents. Although queries are not generally as long as EHR notes, an average patient without adequate medical knowledge may have difficulties constructing effective queries. Thus, this baseline can be considered as a proxy to how a patient actually conducts his own search in the real word.

3.2.2 LDA Models

We trained LDA topic models from over 6000 de-identified EHR notes to identify prominent topics from the test notes. The number of topics are selected based on retrieval performance. LDA models extract distributions over individual word tokens for each topic. However, medical con-

Document Type	Documents (Tokens)	Average Tokens (StdDev)
Health Topics	956 (141,185)	147.7 (37.6)
Medical Encyclopedia	7078 (5,126,101)	724.2 (363.7)
Drugs, Supplements, and Herbal Information	1332 (1,726,570)	1296.2 (992.8)
Total	9366 (6,993,856)	749.1 (565.9)

Table 1: MedlinePlus Collection

cepts often contain more than one token. We employ turbo topics (Blei and Lafferty, 2009) to find phrases from these topics. This method builds significant n-grams based on a language model of arbitrary length expressions. Queries are then generated from the n-grams. We take the top 5 phrases as queries from the topics that has a combined probability of over 80%.

3.2.3 Learning-Based Key Concept Identification

We also developed learning-based key concept identification to build queries from EHR notes. We employed Conditional Random Fields (CRF) model (Lafferty et al., 2001) to identify key concepts, which are most in need of explanation by external education materials. These key concepts can be considered in a broad sense topics, as they also capture various aspects of the EHR note content. We explored lexical, morphological, UMLS (Bodenreider, 2004) semantic type, and word embeddings as features. The word embeddings were induced from a combination of Wikipedia articles in the Medicine category and de-identified EHR progress notes.

Three CRF models are learned using different training data. One uses Wikipedia articles. Wikipedia, especially the Medicine category, is an appealing resource for such information as the human curated links in them are naturally concepts that are important. However, the number of articles in the Medicine category outnumbers our EHR notes substantially, we therefore restricted the Wikipedia articles to the Diabetes category. The paragraphs before the table of contents box are used. Anchor texts in these paragraphs are treated as key concepts. The second model uses EHR notes. Training data for the model was generated from the retrieval gold standard. A phrase in an EHR note would be annotated as a key concept if it matches the title of one of the note's relevant *MedlinePlus* documents. Lastly, we augmented the EHR corpus with Wikipedia articles, and compared its performance to the other CRF-based models. Leave-one-out cross validation is used in the last two models.

	System	P@10	MAP	Increase
1	Baseline	0%	0.0091	-
2	CHV	5%	0.0240	2.6
3	LDA	10%	0.0489	5.4
4	Key (Wiki)	16%	0.0851	9.4
5	Key (EHR)	16.5%	0.0879	9.7
6	Key (Wiki+EHR)	18%	0.1030	11.3

Table 3: System Performance

3.2.4 Query Expansion

We explored query expansion by incorporating relevance feedback from pseudo-relevant documents. The initial queries are generated using methods described previously. Among the top 20 retrieval results, those with a title that matches one of the identified key concepts are considered pseudo-relevant documents. This additional requirement is to ensure that the expanded concepts do not drift from the main topic of the medical notes. From these documents, medical concepts are extracted using MetaMap (Aronson and Lang, 2010). These concepts, with their synonyms provided by UMLS, are used as expansions.

4 Experiment Results

4.1 LDA Models

100 topics are learned from the de-identified EHR note collection. This level of topic granularity shows the best performance in our experiments. The retrieval result is shown in Table 4, row 3. The improvement over the baseline is 5.4 folds, and is statistically significant using a paired Student's t-test ($p < 0.05$).

4.2 Key Concept Identification

The key concept identification performance of the three CRF models is shown in Table 4.2. Retrieval

	Training Data		
	Wiki	EHR	Wiki+EHR
Precision	16.27%	35.92%	33.79%
Recall	26.88%	34.09%	33.18%
F1	18.74%	33.70%	32.54%

Table 4: Key concept identification performance

Training Data	P@10	MAP
Wiki	20% (16%)	0.1067 (0.0851)
EHR	16% (16.5%)	0.0951 (0.0879)
Wiki+EHR	21% (18%)	0.1169 (0.1030)

Table 5: System performance with pseudo-relevance feedback. Numbers in parentheses are without pseudo-relevance feedback.

performance of these models are shown in Table 4, rows 4 to 6. All systems showed a statistical significant improvement over the baseline. The last model's improvement is also statistically significant over the LDA approach. Query expansion methods further improved system performance, as shown in Table 4.2.

5 Discussions

From the LDA model, Table 5 shows the top 10 n-grams from 7 topics trained on the medical text. It is clear that while topics like the first one capture medical concepts, others like the second one do not. The LDA results also highlight the noisy nature of the EHR notes. Queries formed by including the generic or noisy terms such as "continue on" will not benefit retrieval results. Examining the retrieval results, we found that when the prominent topics include medical concepts, the top 10 results usually contain at least one relevant document. When only generic topics are identified, relevant documents are absent in the top 10 results.

The CRF models, on the other hand, are better at capturing the concepts in the EHR notes. All three models outperformed the LDA model. One drawback of using models trained from title-matching phrases is that the identified key concepts fail to cover the scope of the medical concepts contained in an EHR note. On average only 23.6% of the concepts are annotated as key phrases in each note. In the evaluation of the CRF models for their key concept identification performance, inclusion of Wikipedia data slightly decreased the F1 score when they were tested on EHR notes. This can be attributed to the fact that the Wikipedia articles outnumbered the EHR data by 7 times. However, this data helped improve the coverage of the key concepts. In one document, the augmented model identified two more out of the total eight key concepts.

	Phrases with the highest probability
1	dialysis, hemodialysis, catheter, renal failure, renal, coumadin, line, picc line, dialysis catheter, failure
2	job id, today, point, continue on, reasonable, try to, continue, yesterday, left, right
3	continue, patient, job id, pain, patient has, normal, patient s, white count, secondary to, culture
4	liver, ascites, normal, tenderness, fluid, stable, elevated, today, edema, chest
5	preliminary, patient, patient s, time, blood, mmoll, patient has, routine, high, vial
6	diarrhea, abdominal, flagyl, stool, abdominal pain, colitis, abdomen, difficile, fluid, distended
7	bipap, pneumonia, year old, respiratory failure, failure, minutes, requiring, encephalopathy, ards, patient

Table 6: Top 10 n-grams from 7 topics using the LDA model

Finally, to investigate the gap between medical language and lay language, we substituted the medical concepts recognized by MetaMap with their consumer-oriented counterparts created by the Consumer Health Vocabulary (CHV) Initiative (Zeng and Tse, 2006). Performance using the substituted EHR notes as shown in Table 4, row 2 more than doubled. The gap highlights the issue that patients may have difficulty finding relevant health information without assistance due to vocabulary mismatch.

6 Conclusions, Limitations and Future Plan

We have shown that using full EHR notes is ineffective at retrieving relevant education materials. Identifying key concepts of an EHR note and then using the key concepts as query terms re-

sult in significantly improved performance (over 10-fold). Furthermore, a query expansion approach in which key concepts are complemented by other medical concepts from pseudo-relevant documents further improves the performance.

One limitation of our design is that only one physician provided relevancy judgments. Additional annotators would provide a more rigorous set of gold standard, allowing us to measure inter-annotator agreement.

There are several directions we can explore in our future research. Firstly, our key concept identification methods are not optimized for the retrieval results, but for the identification subtask only. We hypothesize that directly optimizing the key concept identifier for retrieval would lead to better performance. We would also investigate domain adaptation techniques to learn key concept identification models from other data sources.

Acknowledgments

This work was supported in part by the Award 1I01HX001457 from the United States Department of Veterans Affairs Health Services Research and Development Program Investigator Initiated Research. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

Alan R Aronson and Franois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–236, June.

Niranjan Balasubramanian, James Allan, and W. Bruce Croft. 2007. A comparison of sentence retrieval techniques. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 813–814, New York, NY, USA. ACM.

Michael Bendersky and W. Bruce Croft. 2008. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 491–498, New York, NY, USA. ACM.

Michael Bendersky, W. Bruce Croft, and David A. Smith. 2009. Two-stage query segmentation for information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 810–811, New York, NY, USA. ACM.

David M. Blei and John D. Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv:0907.1013 [stat]*, July. arXiv: 0907.1013.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–270, January.

Marc-Allen Cartright, Henry A. Feild, and James Allan. 2011. Evidence finding using a collection of books. In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, BooksOnline '11, pages 11–18, New York, NY, USA. ACM.

Tom Delbanco, Jan Walker, Sigall K. Bell, Jonathan D. Darer, Joann G. Elmore, Nadine Farag, Henry J. Feldman, Roanne Mejilla, Long Ngo, James D. Ralston, Stephen E. Ross, Neha Trivedi, Elisabeth Vodicka, and Suzanne G. Leveille. 2012. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Ann. Intern. Med.*, 157(7):461–470, October.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Introduction to the special issue on patent processing. *Information Processing & Management*, 43(5):1149–1153, September.

Debasis Ganguly, Johannes Leveling, Walid Magdy, and Gareth J.F. Jones. 2011. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1953–1956, New York, NY, USA. ACM.

Giridhar Kumaran and Vitor R. Carvalho. 2009. Reducing long queries using query quality predictors. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 564–571, New York, NY, USA. ACM.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chia-Jung Lee and W. Bruce Croft. 2012. Generating queries from user-selected text. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 100–109, New York, NY, USA. ACM.

Parvaz Mahdabi, Linda Andersson, Mostafa Keikha, and Fabio Crestani. 2012. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 505–514, New York, NY, USA. ACM.

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 472–479, New York, NY, USA. ACM.

Donald Metzler, Trevor Strohman, Yun Zhou, and W. B. Croft. 2004. Indri at TREC 2005: Terabyte track.

Matthew Mossanen, Lawrence D. True, Jonathan L. Wright, Funda Vakar-Lopez, Danielle Lavallee, and John L. Gore. 2014. Surgical pathology and the patient: a systematic review evaluating the primary audience of pathology reports. *Hum. Pathol.*, July.

Howard Turtle and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3):187–222, July.

David Wiljer, Sima Bogomilsky, Pamela Catton, Cindy Murray, Janice Stewart, and Mark Minden. 2006. Getting results for hematology patients through access to the electronic health record. *Can Oncol Nurs J*, 16(3):154–164.

Xiaobing Xue and W. Bruce Croft. 2009. Transforming patents into prior-art queries. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 808–809, New York, NY, USA. ACM.

Qing T. Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*, 13(1):24–29, February.