Incorporating Statistical Topic Models in the Retrieval of Healthcare Documents

Karla Caballero[1] and Ram Akella[1,2]

[1]University of California Santa Cruz, 1156 High Street Santa Cruz CA, USA
[2]University of California Berkeley, 94720 Berkeley CA, USA

Abstract

Patients often search for information on the web about treatments and diseases after they are discharged from the hospital. However, searching for medical information on the web poses challenges due to related terms and synonymy for the same disease and treatment. In this paper, we present a method that combines Statistical Topics Models, Language Models and Natural Language Processing to retrieve healthcare related documents. In addition, we test if the incorporation of terms extracted from the patient's discharge summary improves the retrieval performance. We show that the proposed framework outperformed the winner of the retrieval CLEF eHealth 2013 challenge by 68% *in the MAP measure (* 0.5226 *vs* 0.3108*), and by* 13% *in NDCG (*0.5202 *vs* 0.3637*). Compared with standard language models, we obtain an improvement of* 92% *in MAP (*0.2666*) and* 45% *in NDCG. (*0.3637*)*

Introduction

Today an increasing number of healthcare related documents are available on the web. Widely known pages such as *WebMd* and *Mayo Clinic* describe detailed symptoms and treatments for different diseases. Patients often search for treatments and disease care options that are available after they are discharged from the hospital. However, the search in this domain poses challenges such as synonyms and related terms for the same disease and treatment that prevent traditional keyword search from being effective. To address these challenges, an ontology tree is often used to expand the query and to include related terms in the search[1]. In this framework, the documents are decomposed into concepts[2]. However, performing this task with web data is unfeasible due to the high computational cost of the annotation process and the heterogeneity of the corpus when a fixed ontology is used.

We propose a method based on Statistical Topic models to provide global information about documents. These methods provide an unsupervised clustering framework which enables us to increase the weight of related words without including them explicitly in the document or in the query . In addition, we describe a method to incorporate noun phrases in the vocabulary without annotating the entire corpus using Natural Language Processing tools. In this paper, we explore the effectiveness of using patient discharge summaries to provide relevant context to the query and to improve the retrieval performance. Here, we exploit the fact that patients read their discharge summaries and pose related queries afterwards.

Background

Traditionally healthcare document retrieval relies on concepts such as symptoms, medications and diseases. Current approaches are generally based on concept and ontology based information retrieval[2]. These methods often require the documents to be annotated using a list of conceptual terms from ontologies trees or taxonomies[1,2]. Then string matching and rule based retrieval techniques are employed to rank documents. To improve the document score given a query, bag-of-words retrieval methods in the concept domain have been proposed. Castells et. al.[1] introduced a vector space model in ontology based retrieval as a form to improve the precision. Other approaches combine the ontology and the bag-of-words based scoring function to discriminate between concepts[2]. The main drawback of these methods is the annotation task. This process requires documents to be annotated using a fixed taxonomy, which is not feasible when using multi-source web documents due to the discrepancy between the scientific nature of the ontologies and the common language of the web documents, and the size of the corpus. In addition, the annotated text does not reflect

the intensity of a concept inside the document and often requires disambiguation[3]. To alleviate all these issues, the query is often expanded using related concepts obtained from the ontology tree instead of annotating the document[4,5]. Other approaches use Relevance Feedback based methods to disambiguate different concepts inside the document[2]. In contrast, we propose an unsupervised method that exploits the clustering power of Statistical Topic Modeling methods based on bag-of-words. This technique uses the term co-occurrence to establish semantic relations to add context to the query. We also explore if the use of contextual text information extracted from patient's discharge summaries improves the retrieval performance.

Methods

In this section, we address the main components of the methodology: statistical topic modeling, noun phrase extraction, the query expansion using discharge summaries, and the incorporation of all these information sources in the document retrieval based on language models methods.

Statistical Topic Modeling

Statistical Topic Modeling is an unsupervised learning technique that allows us to extract latent topics from a document corpus. The main idea is that the extracted topics provide a global context, which cannot be obtained by using independent word counts as in the standard bag-of-words models. For the current problem, we use the Generalized Latent Dirichlet Allocation (GD-LDA) topic model[6]. This technique has been shown to improve the document retrieval by constraining the number of covariances among the topics. In this model, each document is defined as a statistical mixture of topics, and each topic is defined as a mixture of words. The most dominant topics in the document are those with the highest probability. Similarly, the higher the probability of a given term the more relevant the term is to the latent topic. For model fitting, we follow a Gibbs sampling based approach as detailed in[6]. We estimate the prior distribution parameters that optimize the augmented likelihood which is defined as follows:

$$p(w, z|\alpha, \beta, \gamma) = p(w|z, \beta)p(z|\alpha, \gamma) = \prod_{i=1}^{N_j} \prod_{j=1}^{J} \prod_{k=1}^{K} \phi_{k,i} \times \theta_{j,k} \tag{1}$$

where w is the set of document words in the corpus, z are the topic labels for each word, α and γ are the prior distribution parameters of the topic mixture. Similarly, β is the parameter vector of the prior distribution of word terms in the vocabulary given a topic. θ_j is the topic mixture of the document j and ϕ is the word mixture for each latent topic k. In this framework we need to fix the number of topics K. We fit the model using 10% of the documents in the corpus selected randomly. Once we estimate the prior distribution parameters, we extract the topic content for the remaining documents by sampling the labels for each word in the document given estimated the prior distribution of the word mixture $\hat{\phi}$. Our goal is to depict a real scenario where only a subset of the corpus is used to train the model. We estimate the topic content $\hat{\theta}_j$ for each document using the following equation:

$$\hat{\theta}_{jk} = \begin{cases} \dfrac{\hat{\alpha_k} + \overline{N}_{j,k}}{\hat{\alpha}_k + \hat{\beta}_k + \sum_{l=1}^{K} \overline{N}_{j,l}}, & \text{if } k = 1 \\ \dfrac{\hat{\alpha}_k + \overline{N}_{j,k}}{\hat{\alpha}_k + \hat{\beta}_k + \sum_{l=k}^{K} \overline{N}_{j,l}} \times \prod_{m=1}^{k-1} \dfrac{\hat{\beta}_m + \sum_{l=m+1}^{K} \overline{N}_{j,l}}{\hat{\alpha}_m + \hat{\beta}_m + \sum_{l=m}^{K} \overline{N}_{j,l}} & \text{if } 1 < k < K \\ \prod_{m=1}^{K-1} \dfrac{\hat{\beta}_m + \sum_{l=m+1}^{K} \overline{N}_{j,l}}{\hat{\alpha}_m + \hat{\beta}_m + \sum_{l=m}^{K} \overline{N}_{j,l}} & \text{if } k = K \end{cases} \tag{2}$$

for the topics $k = 1 \dots K$ and the documents $j = 1 \dots J$. The value of $\overline{N_{j,k}}$ is the topic label counts obtained from Gibbs sampling. We assign a topic label z to each observed word inside the document using the following expression:

$$p(z_{wj} = k|z^{\neg wj}, \alpha, \beta, \gamma) = \hat{\phi_{w,j}} \times \hat{\theta_{j,k}^{\neg wj}} \tag{3}$$

where $\neg wj$ represent the estimation without the current word w in document j. We calculate the topic mixture for each document and save it as meta data to be used later in the retrieval task.

Noun Phrases Extraction

Noun phrases provide relevant information about diseases and treatments that are not accounted for when we use the bag-of-words scheme. Several diseases and medications are often identified by two or more terms. There are also several word combinations that can be considered as stop words (i.e., patient name, physician name) that are removed when we index the corpus. However, annotating all the corpus is not a feasible task due to the heterogeneity and size of the corpus. In addition, when we annotate a document, several concept candidates can match the selected text. This fact challenges the performance of entity and concept based retrieval due to the ambiguity of the resulting annotations. To alleviate this problem, we extract the most common noun phrases that depict healthcare related content by annotating a small part of the corpus using the MetaMap[7] extraction tool. This tool uses Natural Language Processing techniques and the UMLS ontology tree[8] to annotate documents. We set up Metamap to return candidates with drug, disease and procedure information by observing their semantic type with the acronym disambiguation option.
After completing the annotation process, we select the noun phrases formed by 2 or more terms which have healthcare related content. Then, we construct a unique identifier that replaces each of the extracted noun phrases inside the document. With this process, we include a typical set of nouns with clinical content used in the web pages without annotating the entire corpus.

Discharge Summaries

Discharge summaries provide additional information to discover the context of the user who poses the query. However, only a fraction of the discharge summary is related to the query. These documents are often formed by a collection of unrelated events. In order to effectively add context to the query, we only take into account those paragraphs with significant relation to the query. We determine this relationship in a unsupervised manner by comparing the query terms and paragraph term distributions using KL divergence and selecting the ones with the smallest distance. We found that in average only 10% of the discharge summary is related to the query. This amount is represented by 1 to 3 paragraphs, compared to an average discharge summary of 10 to 20 paragraphs. Therefore, including the whole summary to expand the query can potentially result in a noisy query. Once we extract the related paragraphs from the discharge summaries, we remove the stop words and terms with low tf-idf. Then, we expand the original query with the processed text.

Retrieval Method

We include the topic information by adding $P_{glda}(q_i|D,\hat{\theta}_j,\hat{\phi})$ to the standard language model defined as follows[9]:

$$\begin{aligned} P(q_i|D) &= \lambda\left(\frac{N_d}{\mu+N_d}\frac{c(q_i|D)}{|D|}+(1-\frac{N_d}{\mu+N_d})P(q_i|C)\right)+(1-\lambda)\left(P_{glda}(q_i|D,\hat{\theta}_j,\hat{\phi})\right) \\ P_{glda}(q_i|D,\theta_j,\phi) &= \sum_{z=1}^{K} p(w|z,\hat{\phi})p(z|\hat{\theta}_j) \end{aligned} \tag{4}$$

where q_i is the vector of query terms, K is the number of topics, $\hat{\phi}$ is the posterior probability estimate of the word mixture for each topic, and $\hat{\theta}_j$ is the topic mixture for the document. We set the smoothing parameter $\mu = 1$ and the combining parameter $\lambda = 0.6$. $N_d = J$ is the number of documents in the corpus.

Results

Experimental Settings

In order to test the proposed framework, we use the corpus of medical-related documents provided by the Khresmoi project[10]. This collection consists of crawled web pages from health and medicine websites that have been certified by the Health on the Net (HON) Foundation as well as other commonly used healthcare websites. The sites cover a broad range of health topics and they target both the general public and healthcare professionals. This dataset consists of $1,628,500$ documents. Due to the nature of the collection, the number of words in each document ranges from 20 to 2000 words. To extract the most common noun phrases, we annotate 15000 documents (1% of the corpus) selected randomly. We extract 6883 noun phrases with two or more terms.
We process the corpus in the following manner: First we extract the text content from the document by removing the HTML tags and headers. Then, we remove special, foreign characters and numbers. Subsequently, we replace the noun phrases by a unique identifier, perform stemming and remove stop words.

Table 1: Document features: Single Words and with noun Phrases extracted

Feature	Single words	Noun phrases
Vocabulary size	$98,734$	$101,497$
Average unique terms per document	767.553	760.522

We use the queries from CLEF eHealth Challenge Task 3 [10], which are formed by 50 medical related queries with a brief description of the user who poses the query (the patient himself, a patient's family or a nurse). Each of these queries have an associated discharge summary or procedure report from the patient that leads the user to pose the query. These text segments are obtained from the MIMIC II dataset [11] which contains the Electronic Medical Record (EMR) from 33000 patients that entered the ICU.

Validation

We test 6 variants of the model for 3 different number of topics. In three of these variants we use standard bag of words: Topic Model only (TM), Topic Model with Discharge Summaries (TM+DS), and a variant of this method by removing the terms with low tf-idf measure (TM+DSTF) (less than 1). In the other variants, we expand the bag of words models mentioned above with Noun phrases (TM+NPh, TM+NPh+DS, TM+NPh+DSTF) for $K = 75, 100$ and 150 topics. Our goal is to test if the number of topics affects the retrieval performance and find the contribution of each model component. Table 1 shows the statistics of the corpus with bag of words and with noun phrases. Here, we observe that the number of noun phrases does not increase the vocabulary size significantly (around 2.7%).
We compare our method with BM25 with pseudo relevance feedback as baseline and with the results of the winner of the CLEF eHealth Challenge Task 3, which is based on Markov Random Fields and term expansion using the Medical Subject Headings (MESH) [5]. This approach uses ontologies to extract the relationships between query concepts in order to expand the query. In addition, we compare the performance of our method with Standard Language Models. Our goal is to show the effectiveness of Statistical Topic modeling in incorporating global context in the retrieval task compared to detailed query expansion using query relationships based on ontologies .

Experimental Results

Table 2 shows the retrieval results based on: Precision at 5 (P@5) and 10 (P@10), Mean Average Precision (MAP) and the Normalized Discounted Cumulative Gain (NDCG). In addition, we show the number of relevant documents retrieved for the 50 queries as a global measure for recall. CLEF eHealth challenge has a total of 1883 relevant labeled documents. Then, the closer the framework is to this number the better global recall the method has.
We observe that our model is consistently better in the $P@5$, MAP and in the $NDCG$ measures for all the variants tested. Our method outperforms the CLEF eHealth winner by 68% in MAP and by 13% in NDCG. The effectiveness of the Statistical Topic Models in the document retrieval is clearly significant. We observe that the TM variant outperforms the standard Language Model framework by 30% in the P@10, 92% in the MAP and 45% in the NDCG measures. We report that the best performance is achieved when we train the model with 100 topics and when we include the Discharge Summaries. This result shows that discharge summaries provide context to the query and consequently improve the retrieval performance. This is particular evident in ambiguous queries or those which contain acronyms, i.e., Shortness of breath (SOB) vs. crying (sob) . In other queries, the addition of the extra context has little or no effect in the performance due to the amount of information already provided by the query. We find that the use of noun phrases does not improve the overall performance. The main reason is that noun phrases are useful when the user searches for a specific type of disease rather than when he looks for general information.

Discussion

We have presented a method to retrieve relevant healthcare documents using Statistical Topic Modeling methods. We showed that the proposed framework outperformed ontology-based approaches. Despite the richness of knowledge extracted by the ontology based techniques, Statistical Topic Modeling captured the document context more effectively for a large corpus of documents obtained from multiple sources in a unsupervised form. We found that the incorporation of discharge summaries in the query improved the overall retrieval performance. These summaries are particularly useful for the case of queries containing general terms and acronyms. Discharge summaries help us to disambiguate query terms by providing additional information to the query.

Table 2: Mean Performance Results of the base model and the variants of the model for the test set

Model	P@5	P@10	MAP	NDCG@10	Doc. Retrieved
Baseline (BM25)	0.4520	0.4700	0.3043	0.4169	1651
CLEF eHealth winner	0.4960	0.5180	0.3108	0.4665	1673
LM	0.4040	0.4040	0.2666	0.3637	1646
K=75					
TM	0.5183	**0.5270**	0.4828	0.5179	1715
TM+DS	0.5224	0.4960	0.4998	0.5103	1014
TM+DSTF	0.5008	0.5204	0.4958	**0.5303**	1023
TM+NPh	0.4920	0.5040	0.4746	0.5059	1697
TM+ NPh +DS	0.5060	0.4959	0.5041	0.4995	1433
TM +NPh +DSTF	0.5320	0.5102	0.5049	0.5170	958
K=100					
TM	0.5224	0.5122	0.4840	0.5117	1722
TM+DS	**0.5428**	0.4840	0.5107	0.5017	1148
TM+DSTF	0.5200	0.5000	**0.5226**	0.5202	1013
TM+NPh	0.5160	0.5166	0.4849	0.5088	1694
TM+ NPh +DS	0.5160	0.4645	0.4881	0.4776	991
TM +NPh +DSTF	0.4020	0.4022	0.4523	0.4235	827
K=150					
TM	0.5265	0.5224	0.4877	0.5186	1718
TM+DS	0.5326	0.5041	0.5011	0.5155	1508
TM+DSTF	0.5306	0.5163	0.5200	0.5239	1445
TM+NPh	0.4760	0.5020	0.5167	0.4899	1670
TM+ NPh +DS	0.5340	0.4632	0.4762	0.4846	842
TM +NPh +DSTF	0.4709	0.4204	0.4464	0.4325	714

Further work includes performing a Learning to Rank algorithm to combine the scores of the retrieval of the TM and TM+DS variants having the clarity of the query as weighting factor. In addition, we plan to explore the decomposition of the discharge summaries into topics. This decomposition would enable us to cluster effectively relevant documents according to their topic mixture. We also plan to exploit the correlations among topics in the discharge summaries as a method of query disambiguation.

References

1 Castells P, Fernandez M, Vallet D. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering. 2007;19(2):261–272.

2 Meij E, Trieschnigg D, de Rijke M, Kraaij W. Conceptual language models for domain-specific retrieval. Information Processing & Management. 2010;46(4):448 – 469. Semantic Annotations in Information Retrieval.

3 Karimi S, Zobel J, Scholer F. Quantifying the impact of concept recognition on biomedical information retrieval. Information Processing & Management. 2012;48(1):94 – 106.

4 Sondhi P, Sun J, Zhai C, Sorrentino R, Kohn MS. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. JAMIA. 2012;19(5):851–858.

5 Zhu D, Stephen W, James M, Carterette B, Liu H. Using Discharge Summaries to Improve Information Retrieval in Clinical Domain, working notes; 2013.

6 Caballero KL, Barajas J, Akella R. The Generalized Dirichlet Distribution in Enhanced Topic Detection. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM; 2012. p. 773–782.

7 Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010;17(3):229–236. Available from: http://dblp.uni-trier.de/db/journals/jamia/jamia17.html#AronsonL10.

8 Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology; 2004.

9 Wei X, Croft WB. LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th ACM SIGIR Conference; 2006. p. 178–185.

10 Suominen H, Salanterä S. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: CLEF 2013. Lecture Notes in Computer Science (LNCS). Springer; 2013. .

11 Saeed M, Lieu G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology. 2002 Sep;29:641–644.