Incorporating Statistical Topic Models in the retrieval of healthcare documents

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

[1] University of California Santa Cruz
[2] University of California Berkeley

Abstract. We present a framework based on Statistical Topics Models, Language Models, Information Extraction, and Ontology Analysis to retrieve healthcare related documents for the CLEF eHealth 2013 Task 3. In this framework we add global information based on latent topics from the documents to improve the document retrieval. We perform six different experiments which consist of a baseline and six variants of the model. Preliminary results show that the use of Language Models with a bag of words scheme results better estimates. However model tunning in the Topic Based model is required to achieve optimal results.

1 Introduction

In these notes we describe the experimental framework presented in the CLEF eHealth 2013 Task 3 which consist of retrieving relevant health care documents. Our main approach consist of providing a solution that does not require any manual input besides the query.

To achieve this goal, we captured global context from the documents in a unsupervised form using Statistical Topic Models. Our main hypothesis is that semantic content may not be correctly represented with single terms alone. In addition, we present a variant of the model where we extract noun phrases and incorporate them in the query. We claim that using the words independently may change the meaning of the query specially when the user is searching documents from a particular disease that is composed by two or more terms. For instance, *cardiac disease* might not be correctly detected as a noun phrase and the independent term *disease* would largely increase the documents retrieved with other types of diseases.

This paper is divided as follows: Section 2 shows the methodology that we follow and how we obtain the different components of the propose model. In section 3 we describe the steps we follow to pre-process the corpus, and we describe the different experiments that we performed. Finally the results and conclusion are discussed.

2 Methodology

In this section we describe the methodology that we use to perform the retrieval of the documents as well as the different components used to run the experiments,

such as the extraction of topics using Statistical Topic Models and the noun phrases extraction. In addition, we detail the incorporation and processing of the discharge summaries in the model.

2.1 Statistical Topic Modeling

Statistical Topic Models is an unsupervised learning technique that allows us to extract latent topics from a corpus of documents. The fundamental idea is that these topics provide a global context, which can not be achieved using only independent words as in the standard bag-of-words modeling. For the current problem, we use Latent Dirichlet Allocation (LDA) [1]. In this model, each document is defined as a mixture of topics and each topic is defined as a mixture of words with a given probability. The most dominant topics in the document are those with highest estimated probability. Similarly, the higher the probability the more important the word is in the topic. All these probabilities are estimated during the model fitting process, which is usually performed using Markov Chain Monte Carlo (MCMC) techniques such as collapsed Gibbs sampling. In this model, we choose the prior distribution parameters that optimizes the augmented likelihood that is defined as follows:

$$p(w, z|\alpha, \beta) = p(w|z, \beta)p(z|\alpha) \tag{1}$$

Where and w is the word, z is the topic label and α and β are the parameters of the prior distribution for the topic mixture and vocabulary mixture respectively. To fit the LDA model with the corpus of healthcare documents we need to select the number of topics. The two parameter vectors that are fitted in the model are the prior distribution vectors for the probability of topics in a document, and the global probability of words given a topic. With these parameters we obtain the topic mixture for each document and save it as metadata that is used in the document retrieval.

2.2 Noun Phrases Extraction

Noun phrases provide relevant information about diseases and treatments usually. This is because several diseases are often identified by two or more terms. In order to select accurately the noun phrases, we use the CTAKES [4] extraction tool that uses medical domain ontology such as SNOMED [5]. We extract all the nouns from the set of discharge summaries of the MIMIC II [3] data. By using this dataset, we are able to find a typical set of clinical nouns used in discharge summaries.

We select the extracted nouns with two or more words since these nouns are not detected by the bag-of-words model. Note that some of the resulting noun phrases are combinations of two of more shorter phrases. We select only those with the smallest number of terms. Then, we analyze if a long noun phrase can be decomposed as a combination of the shortest noun phrases. If this is the case then we ignore this phrase. If this is not the case, we keep the full noun

phrase. We remove the phrases which do not include any medical content. To determine if the medical content of the phrases, we compare the extracted nouns with the SNOMED ontologies in order to keep the ones inside the ontology. We obtain $3,075$ noun phrases using $10,000$ discharge summaries from patients with different diseases.

2.3 Discharge Summaries

We include the discharge summaries information in three of the six experiments we perform. However, when we reviewed the content of the discharge summaries, some of them have little of not relevance with the query. To overcome this problem. We only take into account those paragraphs with significant relation to the query.

We determine this relationship in a unsupervised form by comparing the query terms with the content of the paragraph. We found that only a small percentage of the discharge summaries contains information related to the query, which ranges from 1 to 3 paragraphs, compared to the average discharge summary length of 10 to 20 paragraphs. In some cases only a sentence in the entire summary has a relationship with the query. The extracted information from these summaries is combined with the orignal query in order to create a single expanded query.

2.4 Retrieval Method

We use Language Models to retrieve the documents. In this framework, we estimate the probability that the query is generated by the document. This probability is represented by the following formula:

$$P(Q|\theta_D) = \prod_{i}^{m} p(q_i|D) \tag{2}$$

where $P(q_i|D)$ is the probability that the query term is generated by the document, and this is defined as:

$$P(q_i|D) = \frac{c(q_i, D)}{|D|} \tag{3}$$

here $c(q_i, D)$ is the frequency of the query word in the document and $|D|$ is the total number of terms in the document. In this method, we might find a zero probability for $P(q_i|D)$ when we have a term in the query that is not inside the document. To overcome this problem, we incorporate a smooth factor in the probability $P(q_i|D)$ that is defined as follows:

$$P(q_i|D) = \lambda \frac{c(q_i|D)}{|D|} + (1 - \lambda) P(q_i|C) \tag{4}$$

where $P(q_i|C)$ is the probability that the query word is in the collection. The value of λ ranges from $[0-1]$. For the experiments, we define this value to be 0.5.

We include the topic information by using the framework introduced in [7], which consist of adding $P_{lda}(q_i|D,\theta_D,\phi)$ to the existing language model. This addition is defined as follows:

$$P_{lda}(q_i|D,\theta_D,\phi) = \sum_{z=1}^{K} p(w|z,\phi)p(z|\theta,D) \tag{5}$$

where K is the number of topics, ϕ is the posterior probability estimate of the word mixture for each topic, and θ is the topic mixture for the document. Then the language model is defined as follows:

$$P(q_i|D) = \lambda(\frac{N_d}{\mu+N_d}\frac{c(q_i|D)}{|D|}+(1-\frac{N_d}{\mu+N_d})P(q_i|C))+(1-\lambda)(P_{lda}(q_i|D,\theta_D,\phi)) \tag{6}$$

The value of μ is defined as 1 and N_d is the number of documents. The value of λ is defined to be 0.6. The noun phrases are included in the model as any other word. Consequently, the frequency of those phrases in a document is used to represent the documents.

3 Experimental Settings

3.1 Dataset Preprocessing

We use the corpus of medical-related documents provided by the Khresmoi project. Details of the dataset are found in [6]. We pre-processed the documents in the following manner: First we extract the text from the document by removing the html tags and headers. Then, we remove special and foreign characters by converting all the characters in UTF8 format. With this step we ensure that all the words are in English format. Subsequently we remove all the numbers from the document.

After the data has been cleaned, we indexed all the documents using Indri index engine from Lemur Project [2]. Here we perform stemming of the words using the Krovetz stemmer. In addition, common words are removed in the indexing step.

We have 1628500 documents and a vocabulary size of 98031 after removing documents with less than 20 terms, removing stop words and stemming when we use the bag of words approach. When we use the noun phrases, we replace the noun phrase found in the document with a term code used as a word, and then we index the documents. With this variant we have a vocabulary size of 99538 and 1628500 documents

Once we have indexed all the data, we fit the model using 10% of the documents in the corpus and 75 topics. We select those documents randomly. This

allow us to model the corpus more accurate. 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 using Gibbs Sampling and the prior distribution of the word mixture. Once we have estimated the label for each word, we calculate the topic mixture for each document and save it a metadata in the document. Therefore we have two sets of topic mixtures: one only taking into account the single terms and other that takes into account the noun phrases.

We analyze the queries by parsing them using the constructed index and we remove the stop words. If we use noun phrases, we first replace them with a term code and then we parsed with the index from the document and remove the stop words.

3.2 Description of the Experiments

We submit six different runnings to the task: one baseline and five variants of the model. In this section we describe each of the variants of the model.

We use as baseline the language models described in equation 4 with only the single terms, and without any noun phrases. This method is considered to be the baseline for the other variants of the model.

Variants of the model

We have 5 variants of the model:

1. Including the information from the Discharge summaries and noun phrases using Probabilistic Topic Modeling and Language Models defined in equation 6.
2. We use noun phrases with the queries that contain the information of the Discharge summaries and using only Language Models
3. Language Models and the information extracted from Discharge Summaries without any noun phrases
4. Language models with noun phrases and topic information using only the queries
5. Language models with noun phrases without any information of the discharge summaries

The main idea is to test how the different components contribute to the global result. All the variants are fast in the retrieval process. The main constrain is the offline process that they require.

4 Results

In this section we will discuss the results of all the variants of the model that we run. Table 1 shows the results of the Precision at 10 (P@10), Mean Average Precision (MAP) and the Normalised Discounted Cumulative Gain (NDCG) for the baseline and the five variants of the model

Table 1. Mean Performance Results of the base model ant the variants of the model for the test set

Model	P@10	MAP	NDCG
Baseline	0.4040	0.2666	0.3637
Variant 1	0.0600	0.0178	0.0548
Variant 2	0.1920	0.1590	0.1765
Variant 3	0.2320	0.1634	0.2062
Variant 4	0.0580	0.0197	0.0549
Variant 5	0.3640	0.2270	0.3281

As we can observe the baseline results in better results than the variants. This can be explained by the nature of the query. In this case the use of noun phrases (Variant 5) may help in some queries but in others affect the result. We observe that the use of Topics does not improve the performance of the method, which may be cause because the topics may add noise to the query instead of clarifying the query. This phenomena could be caused by the number of topics, which are not sufficient to describe correctly the corpus. Other explanation is that some results retrieved by the method may not be labeled. This may result in a poor performance of the methods.

Figure 1 shows the precision results for all the queries using the best method (baseline). Here we can observe that the precision of some queries is negative. This means that the baseline have lower performance that the median of the other groups. However some other queries has a comparable performance of the best groups.

5 Conclusion

In this paper we presented a method to retrieve relevant health care related document. We believe that our approach have promising results. Further research and model tunning may be required in order to make our approach competitive and suitable to be applied in the healthcare context.

References

1. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022, 2003.
2. University of Massachusetts and Carnegie Mellon University. The lemur project, 2010.
3. M. Saeed, G. Lieu, and R. G. Mark. MIMIC II: a massive temporal icu patient database to support research in intelligent patient monitoring. *Computers in Cardiology*, 29:641–644, September 2002.
4. Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text

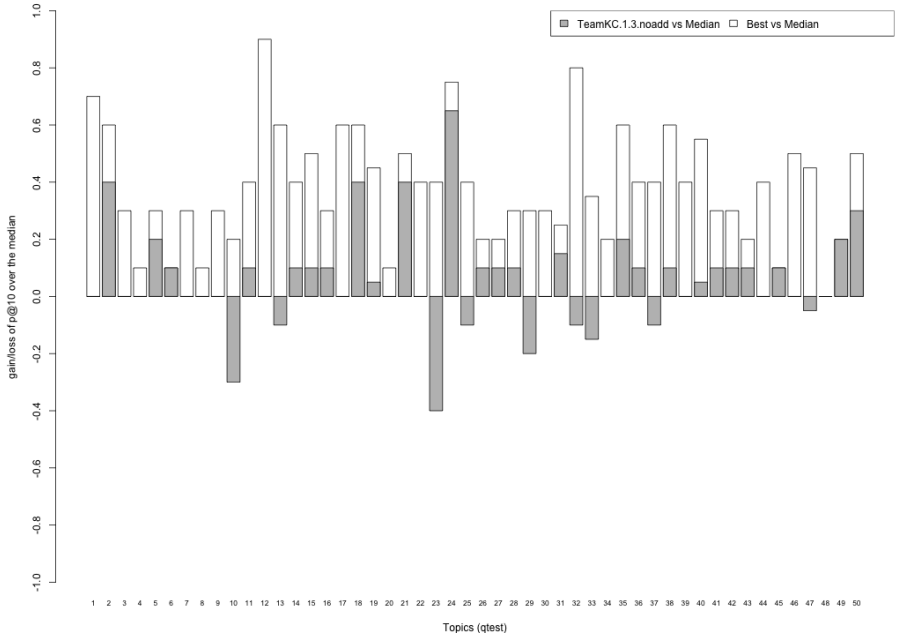

Fig. 1. Precision at 10 for each query inside the test set

analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, September 2010.

5. Kent A. Spackman, Ph. D, Keith E. Campbell, Ph. D, Roger A. Cote, and D. Sc. (hon). SNOMED RT: A reference terminology for health care. In *J. of the American Medical Informatics Association*, pages 640–644, 1997.
6. Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Danielle Mowery, Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *CLEF 2013*, Lecture Notes in Computer Science (LNCS). Springer, 2013.
7. Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.