NOVASearch at Precision Medicine 2017

Gonçalo Araújo, André Mourão, João Magalhães
NOVA Laboratory for Computer Science and Informatics
Faculdade de Ciências e Tecnologia
Universidade NOVA de Lisboa
2825-516 Caparica
{gc.araujo, a.mourao}@campus.fct.unl.pt, jmag@fct.unl.pt

ABSTRACT

This article describes the NOVASearch retrieval system for TREC 2017 Precision Medicine Track. The parsing of queries and documents in the Clinical Trials task were structured into multiple fields concerning the detail about inclusion and exclusion criteria for the trials. We also considered multiple text processing filters on the largest text fields. We also implemented a pseudo relevance feedback(PRF) query expansion, incrementally queering the data set and creating new queries, using important terms of the best ranked documents, previously retrieved.

Keywords

Medical Text Retrieval; Query expansion; Information Retrieval

1. INTRODUCTION

The TREC Precision Medicine Track 2017, aims to provide clinicians with important information to support medical decisions. The clinical decision support will be focused on a specific use case, cancer patients, so that clinicians can have access to very specific medical trials. This year the TREC challenge will be divided into two goals(each one containing a specific data set):

- **Biomedical Articles:** Retrieval of articles that reefer to interventions made by other doctors, to patients suffering from the same cancer as the patients being treated.
- **Clinical Trials:** Retrieval of clinical trials were the patient could be a participant due to his disease.

Section 2 details the indexing and retrieval methods implemented in both tasks for the TREC 2017 Topics (queries). Section 3 explains the usage of query expansion. Section 4 discusses the evaluation results.

2. METHODS AND ALGORITHMS

Both indexing and retrieval methods were implemented with Apache Lucene, a text search engine library for Java, that contains very helpful methods for creation of indexes, Queries and text search.

2.1 Documents parser

The documents parser uses a pipeline of filters so that we can mitigate factors like, the number of irrelevant words and variation of words (ex.reduction to primal verbal form *driving* to *drive*), which leads to more effective retrieval results. Most of the methods are widely used in Information Retrieval, that were tuned to the tasks at hand:

- **Tokeinzation:** Used to remove all form of punctuation and split text into tokens. We also converted all words into lower case.
- **Stop word removal:** Remove specific words like "this", "a", "or", that will occur in most of the English texts.
- **Word grams:** This filter creates tokens from other tokens. We tested a range of minimum and maximum size of neighboring words to create the indexing tokens. For example, considering a minimum of 2 and a maximum of 3, we get the following word grams for the sentence *"Words have no meaning"*:

```
Words have
Words have no
have no
have no meaning
no meaning
```

- **Stemming:** We used the Snowball filter to stem the indexed documents. Stemming is the process of reducing words to their word stem. For example: *cooking*, *cooker*, and *cooks* would all be reduced to their root word *cook*.
- **Character grams:** Creates n-grams of words, with a minimum and a maximum length for the words is given as a parameter. This is a technique that has been been successful in the medical domain due to the complex spelling of medical terms (many common prefixes and suffixes). For example, using 3 as a minimum, 5 as maximum, in the sentence *"Good afternoon"*, we would have the following n-grams: `Goo, Good, ood, aft, afte, after, fte, fter, ftern, ter, tern, terno, ern, erno, ernoo, rno, rnoo, rnoon, noo, noon, oon.`
- **Demographics filter** Removes the demographic information from the trials, *table 1* is a good example. Creates age range and a gender exclusion criteria. This type of processing can be found in a similar way in the PICO(population, Intervention, Control, Outcome) fields extraction [7].

2.2 Indexing

The TREC PM 2017 collection of clinical trials, contains a large number of trials and most of them not specific for our topics. To index the information contained in the documents, it was necessary to choose the fields that could be relevant in a medical environment and index them. After some inspection we fields to index *gender*, *minimum age*, *maximum age*, *title*, and *condition(disease)*. The utilization of the patient specific information can be seen in other articles with proven results [2].

The provided format of the documents in the collection, a field-structured document, helped us to minimize the workload on the pre-processing of the text before indexing it. For example, **<minimum_age>** *18* Years **</minimum_age>**, after retrieving the text of the xml tag **<minimum_age>** , there's only need to split the text by white spaces and the result is an array containing [”18”,”years”], we know for sure that the first value on the array is a string containing a number, that represents the minimum age of the patients for this clinical trial.

The information contained in the larger text fields such as, *brief_title* or *summary* are first processed using the analysis process explained in the previous section, and then indexed. Not all the fields are available in all clinical trials, however, in our implementation we stored all the fields even if they're empty, or nonexistent on the clinical trial. The indexed fields are as follows:

- **Text:** Field containing a concatenation of relevant text fields about the intervention.
- **Official Title:** The official title of the clinical trial.
- **Brief Title:** A brief title containing only some keywords of the title.
- **Brief Summary:** A excerpt of the summary.
- **Detailed Description:** An detailed description specifying the intervention made on the clinical trial.
- **Criteria:** Specific Inclusion and Exclusion criteria for the patients (other deceases, allergies, other drugs previously used).
- **Min/Max Age:** The age range for patients acceptable for this trial.
- **Gender:** The genders acceptable for the trial.

In the final submission only the *Text* , *Criteria* and *Min/Max age* and *gender* fields were used for the Clinical trials retrieval.

In the Scientific Abstracts task we indexed the abstract with multiple features and aggregated the corresponding ranks for computing the final search results.

2.3 Retrieval Models

To retrieve relevant documents for each topic (a clinical case), we first did some processing of the TREC topics text fields, corresponding to the same steps we did to the indexed fields. We examined multiple runs using different types of analyzers, so we could test which analyzers would do a better job in filtering MESH terms [1] and overall medical relevant terms. We also used several types of query parsers and various ranking functions. Our main focus was to create the maximum number of inclusion and exclusion criteria, querying the collection of documents with information relevant to the indexed fields, this way we could take advantage of having structured documents in our data set and in our topics.

2.3.1 Vector Space Model

Term Frequency-Inverse Document Frequency consists, in a statistical method to define the importance of words for a specific document and simultaneously for the collection. If a word frequency in a collection is low, and its frequency in a document is high its tf-idf value is higher than the value for words that occur more often in the collection, even if they have high frequency in a few documents.

2.3.2 BM25

Ranking function that uses TF(term frequency) IDF(inverse document frequency) functions to rank documents according to the terms from the query.

$$\sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}}\right)}, \qquad (1)$$

After some tests, we ended up using the BM25 model.

3. QUERY EXPANSION

Our implementation of PRF query expansions is based on Lucene MoreLikeThis (MLT), and previously implemented in [6] by some of the members of the NOVASearch team. PRF is a method for automatic local analysis [3], providing the user with possible relevant terms for a new query, without any work from the user. We expanded our queries using the top terms of the ”*text*” field from the results of queries made to the collection of documents. We used the top-3 retrieved documents, and the following base parameters:

- **Min Document Frequency (Baseline=5):** Minimum number of documents the terms should occur.
- **Min Term Frequency (Baseline=2):** Minimum number of times the term should occur in the documents.
- **Min Word Length (Baseline=3):** Minimum length of the term.
- **Max Query Terms(Baseline=15):** Number of terms extracting from the document to insert in the query.

Our expanded queries end up being a ”bag-of-words”, containing the 15 terms that were relevant in the previously retrieved document. For example if our initial query contains the term *”Adenocarcinoma”* (type of lung cancer), we can create a MLT query with terms like, *”lung cancer”*, *”NSCLC”* (Non-Small Cell Lung Cancer). The final query will contain the initial queries, and 3 new expansions, each created from one of the top-3 Documents.

4. EVALUATION

For this year track we implemented two different types of systems. The first one, for the retrieval of Scientific Abstracts, implemented query expansion with PRF and MESH terms, and also Recripocal Rank Fusion (RRF) using multiple ranking features. Both query expansion and rank fusion were based on [4, 5]. The second system, for the Clinical Trials task, consisted in applying BM25, and Query expansion with PRF and runs with multiple combinations of search results filters based on demographic, and exclusion criteria to reduce non relevant retrievals.

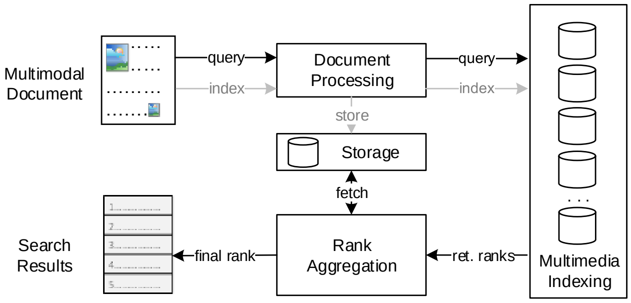

Figure 1: The NOVASearch retrieval system architecture considers multiple features and a rank fusion of the retrieval system.

4.1 Scientific Abstracts

The initial query is the disease, gene variant, patient demographics and existing conditions text. MESH and SNOmed are medical vocabularies used specifically for the retrieval of medical information. In general, the runs for retrieving Scientific Abstracts are:

- **Run 1:** Search (BM25L similarity) in the Medline/ ASCO/ AACR title and abstract text. Query is the disease, gene variant, patient demographics and existing conditions text, expanded by PRF using terms from the top 25 results. Query also expanded with synonyms, alternative and preferred terms from MeSH.
- **Run 2:** Search (multiple similarities) in the Medline/ASCO/AACR title and abstract text. Query is the disease, gene variant, patient demographics and existing conditions text, expanded by PRF using terms from the top 25 results. Query also expanded with synonyms, alternative and preferred terms from MeSH. Final rank is the fusion of runs using BM25L,BM25+,TF-IDF and Dirichlet Language Model similarities, using RRF:

$$RRFscore(d\epsilon D) = \sum_{r\epsilon R} \frac{1}{k + r(d))} \qquad (2)$$

- **Run 3:** Search (multiple similarities) in the Medline/ASCO/AACR title and abstract text. Query is the disease, gene variant, patient demographics and existing conditions text, expanded by PRF using terms from the top 25 results. Query also expanded with synonyms, alternative and preferred terms from MeSH and SNOMed. Final rank is the fusion of runs using BM25L,BM25+,TF-IDF and Dirichlet Language Model similarities using RRF.

4.2 Clinical Trials

The base Query is the disease, gene variant. patient demographics are used as filters to inclusion and exclusion criteria. PRF query expansion using top 3 documents retrieved and exclusion criteria based on other conditions the patient may suffer. In general, the runs for retrieving Clinical Trials are:

Run Id	RRF	PRF	Syns.	MESH	SNOMed
1		x	x	x	
2	x	x	x	x	
3	x	x	x	x	x

Table 1: The different methods used on each Run Id. Title and abstract were indexed from the Scientific Abstracts. These fields are the ones being searched throughout all the runs.

- **Run 1:** Search (BM25 similarity) in the trial title, summary and description text. Query is the disease and gene variant text.
- **Run 2:** Search (BM25 similarity) in the trial title, summary and description text. Query is the disease and gene variant text. Results filtering by the patient age and gender.
- **Run 3:** Search (BM25 similarity) in the trial title, summary and description text. Query is the disease and gene variant text, expanded by PRF using terms from the top 3 results. Results filtering by the patient age and gender.
- **Run 4:** Search (BM25 similarity) in the trial title, summary and description text. Query is the disease and gene variant text. Filtered results by matching the patient age and gender to trial's criteria, and where the patient's existing conditions exclusion criteria matched the trails exclusion criteria.
- **Run 5:** Search (BM25 similarity) in the trial title, summary and description text. Query is the disease and gene variant text, expanded by PRF using terms from the top 3 results. Filtered results by matching the patient age and gender to trial's criteria, and where the patient's existing conditions exclusion criteria matched the trails exclusion criteria.

Run Id	Demographics	PRF	Cond. Exc
1			
2	x		
3	x	x	
4	x		x
5	x	x	x

Table 2: The different methods used on each Run Id. Title, Summary and description were indexed from the Clinical trials. These fields are the ones being searched throughout all the runs.

4.3 Results and Discussion

As we can see by the results in table 3, run 2 and 3 usage of RRF Rank Fusion lead us to an increase in precision and in Discount Cumulative Gain, retrieving more relevant documents than run 1. SNOmed terms were not a good addition to our query, resulting in a worst performance than only using MESH terms.

For the Clinical Trials, PRF query expansion was not a good addition, all the runs using this implementation got the

Run Id	infNDCG	P10	R-prec
1	0.196	0.252	0.137
2	**0.226**	**0.314**	**0.156**
3	0.209	0.29	0.144

Table 3: Results for the Scientific Abstracts Task.

Run Id	P5	P10	P15	Recall	MAP
1	0.421	0.386	0.326	0.733	0.246
2	**0.450**	**0.400**	**0.345**	0.733	**0.253**
3	0.386	0.329	0.312	**0.786**	0.221
4	0.443	0.389	0.343	0.665	0.238
5	0.379	0.325	0.310	0.723	0.210

Table 4: Results for the Clinical Trials Task.

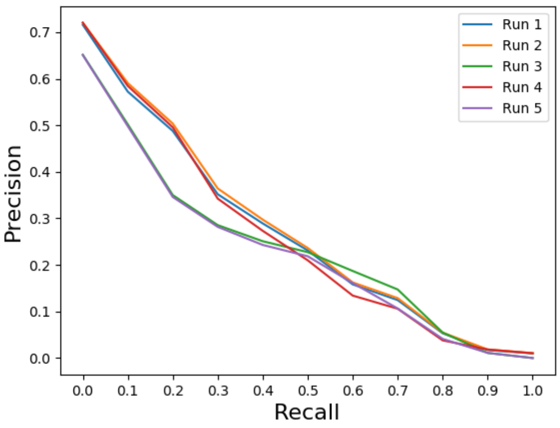

Figure 2: Precision-recall graph illustrates the different runs.

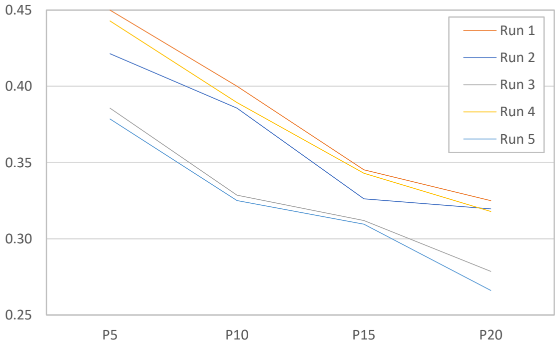

Figure 3: Precision at top retrieved results, i.e., P@5, P@10, P@15 and P@20.

worst performance, while applying simple filtering to create exclusion criteria for the trials improved the performance of the retrieval system. The second run using only Demographic filtering was the best run, followed closely by the run number 4, which added the patient diseases as an exclusion criteria.

5. CONCLUSION

The overall results for the Clinical Trials retrieval were very positive, most of the time above the median. Using query expansion and filters based on the demographic information of the patient proved to be a good solution for the retrieval on both tasks, even if in some cases query expansion was not the overall best. Other good result related to the described runs is the number of topics where we had equal or better results than the median, 22 topics. In 4 of those topics we retrieved documents while the median retrieved 0 documents, showing that our implementation of query expansion can overcome topics where probably no relevant terms for matching were found between trials and topics. Other good indication of our runs are the 5 topics where we matched the best result. As future work, we will process relevance judgment to train the query expansion method.

References

[1] J. Dutkiewicz, C. JÄŹdrzejek, M. FrÄĚckowiak, and P. Werda. Put contribution to trec cds 2016.

[2] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford. Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104(2):444–466, Feb 2016.

[3] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[4] A. Mourão, F. Martins, and J. Magalhães. Novasearch at trec 2015 clinical decision support track.

[5] A. Mourao, F. Martins, and J. Magalhaes. Novasearch at trec 2014 clinical decision support track. Technical report, UNIVERSIDADE NOVA DE LISBOA (PORTUGAL), 2014.

[6] A. Mourão, F. Martins, and J. Magalhães. Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, 39:35–45, 2015.

[7] H. Scells, G. Zuccon, B. Koopman, A. Deacon, L. Azzopardi, and Geva. Integrating the framing of clinical questions via pico into the retrieval of medical literature for systematic reviews.

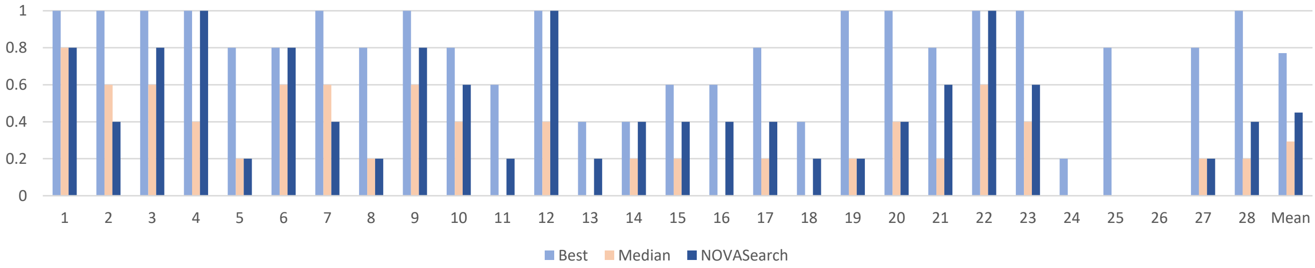

Figure 4: Precision at 10 for all evaluated queries.

Topic	Disease	Gene	Demographic	Other
1	Liposarcoma	CDK4 Amplification	38-year-old male	GERD
2	Colon cancer	KRAS (G13D), BRAF (V600E)	52-year-old male	Type II Diabetes, Hypertension
3	Meningioma	NF2 (K322), AKT1(E17K)	45-year-old female	None
4	Breast cancer	FGFR1 Amplification, PTEN (Q171)	67-year-old female	Depression, Hypertension, Heart Disease
5	Melanoma	BRAF (V600E), CDKN2A Deletion	45-year-old female	None
6	Melanoma	NRAS (Q61K)	55-year-old male	Hypertension
7	Lung cancer	EGFR (L858R)	50-year-old female	Lupus
8	Lung cancer	EML4-ALK Fusion transcript	52-year-old male	Hypertension, Osteoarthritis
9	Gastrointestinal stromal tumor	KIT Exon 9 (A502$_{Y}$503dup)	49-year-old female	None
10	Gastric cancer	PIK3CA (E545K)	54-year-old male	Depression
11	Cholangiocarcinoma	BRCA2	72-year-old male	Diabetes
12	Cholangiocarcinoma	IDH1 (R132H)	64-year-old male	Neuropathy
13	Cervical cancer	STK11	26-year-old female	None
14	Pancreatic cancer	CDKN2A	54-year-old male	Diabetes, Hypertension
15	Prostate cancer	PTEN Inactivating	81-year-old male	Hypertension, Depression
16	Pancreatic cancer	CDK6 Amplification	48-year-old male	None
17	Colorectal cancer	FGFR1 Amplification	35-year-old female	None
18	Liposarcoma	MDM2 Amplification	26-year-old male	None
19	Lung adenocarcinoma	ALK Fusion	64-year-old female	Emphysema
20	Lung cancer	ERBB2 Amplification	70-year-old male	Arthritis
21	Breast cancer	PTEN Loss	54-year-old female	Congestive Heart Failure
22	Lung cancer	NTRK1	58-year-old female	Depression, Hypertension, Diabetes
23	Lung adenocarcinoma	MET Amplification	48-year-old male	Emphysema
24	Breast cancer	NRAS Amplification	35-year-old female	None
25	Pancreatic adenocarcinoma	KRAS, TP53	49-year-old female	None
26	Pancreatic ductal adenocarcinoma	ERBB3	73-year-old female	Whipple, FNA
27	Ampullary carcinoma	KRAS	61-year-old male	None
28	Pancreatic adenocarcinoma	RB1, TP53, KRAS	57-year-old female	None

Table 5: Full list of Clinical Trials query topics: disease, gene, demographic, and other.