The University of Padua IMS Research Group at TREC 2018 Precision Medicine Track

Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin

Department of Information Engineering
University of Padua, Italy
{maristella.agosti, giorgiomaria.dinunzio, stefano.marchesin}@unipd.it

Abstract. We report on the participation of the Information Management System (IMS) Research Group of the University of Padua in the second task of the Precision Medicine Track at TREC 2018: the Clinical Trials task. We designed a procedure to: i) expand query terms iteratively, based on knowledge bases, to increase the probability of finding relevant trials by adding `neoplasm`, `gene`, and `protein` term variants to the initial query; ii) filter out trials based on demographic data. We submitted three runs: a plain BM25 using the provided textual fields `<gene>` and `<disease>` as query, a BM25 with a first knowledge-based query expansion, and another BM25 with an additional knowledge-based query expansion. This initial set of experiments lays the ground for a deeper study on the effectiveness of (automatic) knowledge-based expansion techniques in the context of precision medicine.

Keywords: Precision medicine, query expansion, cancer-gene-protein relationships

1 Introduction

The TREC 2018 Precision Medicine (PM) Track[1] focuses on a relevant use case in clinical decision support: to provide useful precision medicine-related information to clinicians treating cancer patients. Each case of a patient is composed of: the disease (i.e. type of cancer), the genetic variants of the disease (i.e. which genes), and some basic demographic information of the patient (i.e. age, gender). Given the condition of a patient, the track proposes two challenges with two different corpora: 1) retrieve the relevant scientific literature about treatments for the specific condition (*Literature Articles*), 2) find relevant clinical trials for which the patient is eligible (*Clinical Trials*).

In our participation to TREC 2018 PM Track, we focused on the Clinical Trials task. In this task, relevant clinical trials constitute the potential for connecting patients with experimental treatments if existing treatments have been ineffective. In this paper, we present the experiments we carried out using a fully automated system based on a procedure to: i) iteratively expand query terms,

[1] http://www.trec-cds.org/2018.html

relying on medical knowledge bases [2, 3], to increase the probability of finding relevant trials by adding neoplasm, gene, and protein term variants to the initial query; ii) filter out inappropriate retrieved trials relying on demographic data.

The aim of this work is twofold: we want to evaluate how a recall oriented approach based on an increasing (and more aggressive) Query Expansion (QE) method affects precision in this context, and study whether the effectiveness of the retrieval approach can be correlated to the quality of the relations contained within the knowledge base used in the query expansion process. In Section 2, we describe the approach we used to index, retrieve and filter clinical trials; in Section 3, we present the experiments and the results of each run; in Section 4, we give our concluding remarks.

2 Methodology

In this section, we describe in detail the methodology we used to produce the ranked list of clinical trials. In particular, we present: i) what pieces of information were indexed, ii) the ranking model and the query expansion approach, iii) how documents were filtered out based on the eligibility of the patient.

2.1 Indexing Step

We first extracted from each document of the collection all the UMLS[2] concepts, focusing on `gene` and `neoplasm` semantic types,[3] using MetaMap [1], a state-of-the-art biomedical concept mapper developed at the National Library of Medicine (NLM). Then, we indexed the document collection by the following fields: `<docid>`, `<text>`, `<max_age>`, `<min_age>`, `<gender>` and `<concepts>`.

2.2 Retrieval Step

We used BM25 [5] as the core model for our experiments. Given a query Q, containing keywords $q_1, ..., q_n$, the BM25 score of a document D is:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl})}, \quad (1)$$

where $f(q_i, D)$ is the frequency of the i-th query term in the document D, dl is the length of the document D, and $avgdl$ is the average document length in the text collection from which documents are drawn. $IDF(q_i)$ is the Inverse Document Frequency (IDF) weight of the query term q_i, and it is computed as:

$$IDF(q_i) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \quad (2)$$

[2] `https://www.nlm.nih.gov/research/umls/`

[3] `https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml`

where N is the total number of documents in the collection and $n(q_i)$ is the number of documents containing q_i. When performing the retrieval using plain BM25 we consider the original query terms belonging to text fields `<disease>` and `<gene>`.

In the following, we describe how we combined the core model with two different knowledge-based query expansions techniques.

BM25 + a Priori QE: With this approach, we perform a knowledge-based a priori query expansion. First, we extract from each query all the UMLS concept codes belonging to `gene`, `neoplasm` and `protein` semantic types using MetaMap. Second, for each extracted concept, we consider all its name variants. Lastly, the expanded query consists in the union of the original terms and the set of name variants. For example, consider a query that contains the word `melanoma` which is mapped to the UMLS concept code `C0025202`. The set of name variants for `melanoma` contains, among others: cutaneous melanoma; malignant melanoma; melanoma; melanoma malignant; mm - malignant melanoma; malignant melanomas; malignant melanoma (disorder); etc. The query will be expanded with the union of the terms of the initial query and all these terms.

BM25 + a Priori QE + Pseudo Relevance Feedback QE: Based on the ranking obtained with the a priori query expansion, we perform an additional pseudo relevance query expansion by taking the top k returned documents and looking for the extracted UMLS concept codes that are matched. The additional expansion is performed by adding to the expanded query the name variants of the neighbour concepts which present a semantic relation[4] with the extracted concepts within the UMLS Metathesuarus. We restrict this second expansion only to the concepts belonging to `gene`, `neoplasm` and `protein` semantic types.

2.3 Filtering Step

Within each clinical trial, one of the most – if not the most – important sections is the `eligibility` section. The eligibility section comprises, among the others, three important demographic aspects that a patient needs to satisfy to be considered eligible for the trial, namely: `minimum age`, `maximum age` and `gender`; where `minimum age` is the minimum age required for a patient to be considered eligible for the trial, `maximum age` is the maximum age required for a patient to be considered eligible for the trial; `gender` is the required gender for the patient to be considered for the trial.

We filter our the clinical trials for which a patient is not eligible whenever his/her demographic data (i.e. the field `<demographic>`, containing age and gender) do not satisfy the eligibility criteria. In those cases where demographic data

[4] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html`

is not complete, we keep all the clinical trials regardless the value of the missing data. For example, when minimum age is missing, we keep the clinical trial regardless the lower bound on the age.

3 Experiments

We submitted three runs for the Clinical Trials task of the 2018 PM track:

1. TERM_BASED refers to the core BM25 model run;
2. NO_PRF refers to the a priori query expansion run;
3. PRF refers to the a priori plus pseudo relevance feedback query expansion run.

Before analyzing the results, we summarize the procedure used for each run:

- Indexing
  - Extract from each document in the collection all the UMLS concept codes restricted to `gene` and `neoplasm` semantic types using MetaMap;
  - Index documents by the following fields: `<docid>`, `<text>`, `<max_age>`, `<min_age>`, `<gender>`, `<concepts>`;
  - Extract from each query all the UMLS concept codes restricted to `gene`, `neoplasm`, and `protein` semantic types using MetaMap;
  - Obtain for each query all the name variants from the extracted concepts.
- Querying
  - Perform a term-based search using `<gene>` and `<disease>` text fields with plain BM25 (TERM_BASED);
  - Perform a first query expansion by using i) all `gene`, `neoplasm`, and `protein` concepts name variants, ii) `gene` and `neoplasm` extracted concepts (NO_PRF);
  - Based on the previous ranking, take the first k documents (k = 10) and look for the extracted concepts that are matched in the top k clinical trials, then perform an additional query expansion by adding the name variants of the neighbour concepts – restricted to `gene`, `neoplasm`, and `protein` semantic types – which present a semantic relation with the matched concept within UMLS (PRF).
- Filtering
  - Filter out retrieved clinical trials for which the patient is not eligible in terms of age and gender.

We used Whoosh, a pure Python search engine library,[5] for indexing, querying and filtering. We used the default values set by Whoosh for the BM25 parameters that are $k_1 = 1.2$ and $b = 0.75$.

[5] `https://whoosh.readthedocs.io/en/latest/intro.html`

team	run	measure	score
hpi-dhc	hpictall	infNDCG	0.5545
Cat_Garfield	MSIIP TRIAL1	infNDCG	0.5503
ims_unipd	IMS_TERM	infNDCG	0.5395
Cat_Garfield	MSIIP TRIAL1	RPrec	0.4294
ims_unipd	IMS_TERM	RPrec	0.4128
Poznan	BB2_vq_noprf	RPrec	0.4101
Cat_Garfield	MSIIP TRIAL1	P@10	0.6260
ims_unipd	IMS_TERM	P@10	0.5660
Poznan	BB2_vq_noprf	P@10	0.5580

Table 1: A comparison of the top three runs for the Clinical Trials task.

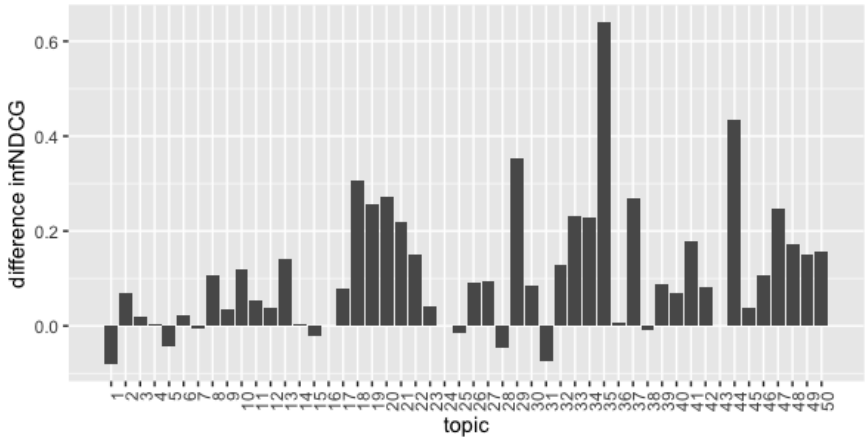
3.1 Results

In this section, we describe the results obtained in this task by reporting the performances published in the overview paper of the 2018 Precision Medicine Track [4] and analyzing for each topic the values of the three performance measures chosen by the organizers: inferred NDCG (infNDCG) [6], precision at 10 (P@10), and R-precision (RPrec).

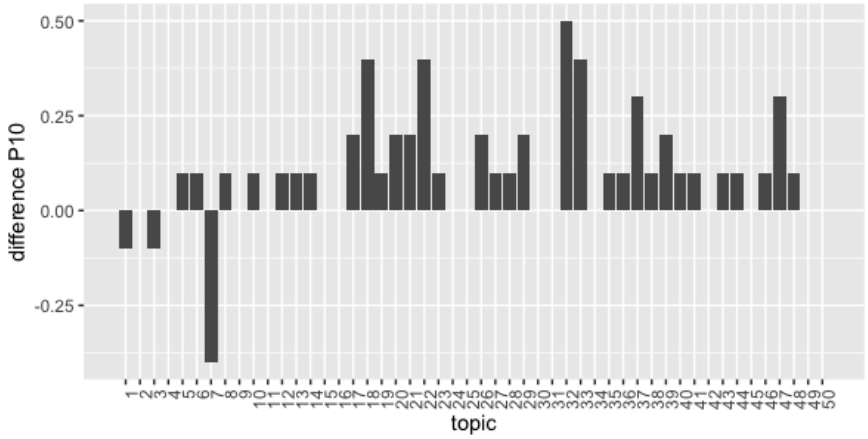
In Table 1, we report the performance of the top three runs for the Clinical Trials task for each measure ([4], Table 6, page 12). The TERM_BASED run is indicated in this table as IMS_TERM. Among all the runs sent by the participants in the task, the core BM25 approach reached the top of the list of the best performing runs. This showed that a plain BM25 with a filtering step of the clinical trials based on the demographic data of the patient, performed as good as other complex approaches.

In order to compare the performance of our three runs and, in particular, to study the impact of the two query expansion approaches, we show the results of the three runs – topic by topic and compared to the median values of the Track – in Table 2, 3, and 4. On average, the TERM_BASED run performed about 10 points percent better compared to the median values for all the measures. The other two runs, performed worse than the median values and, in general, much worse than our baseline.

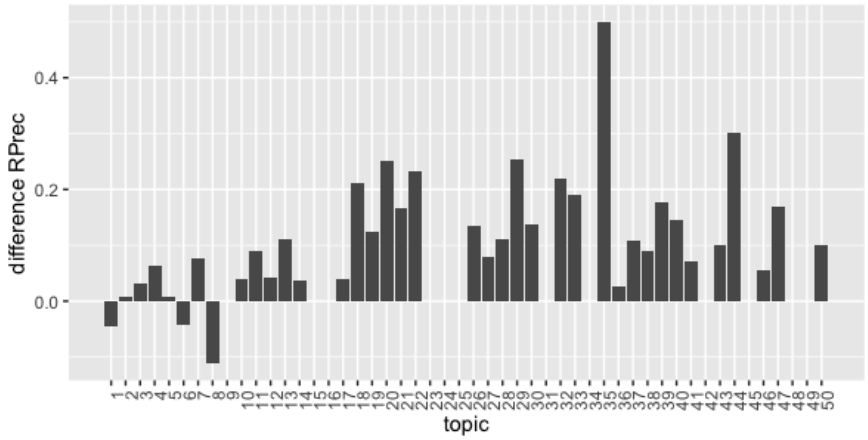
In order to better visualize the comparison between the TERM_BASED run and the median values, we plotted the differences between the two values for each topic. In Figure 1a, 1b, and 1c we highlight the fact that the TERM_BASED run performs consistently better than the median values of the track. It is worth investigating why topic 1 is the only topic where this run performs worse for all the three measures; moreover, topic 4 shows the worst performance for P@10 (considerably large in terms of magnitude). In some cases, for example topic 25 and topic 31, when our run does not show any difference in terms of P@10 and RPrec, the value of infNDCG is slightly worse than the median.



(a) Difference between run and track median values of infNDCG



(b) Difference between run and track median values of P@10



(c) Difference between run and track median values of RPrec

Fig. 1: Difference between TERM_BASED run an median values of the track.

4 Final Remarks

In this paper, we presented the results of our participation in the TREC 2018 Precision Medicine Track. In particular, our objective was the study of a query expansion approach together with a pseudo-relevance feedback one, followed by a filtering phase in the Clinical Trials task.

The analysis of the results shows that our baseline, a plain BM25 with a filtering of the clinical trials based on the demographic data of the patient, performs well compared to all the runs of the task. In particular, the performance of the run was comparable with the top performing runs which are based on more complex retrieval models. On the other hand, the query expansion approach introduces too much noise and decreases the performances in a significant way.

Our next step is to perform a failure analysis to understand whether the use of a specific medical knowledge base instead of a metathesaurus (such as UMLS), and the choice of a more accurate set of relations between terms, can improve the performance of a query expansion approach.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. pp. 17–21. American Medical Informatics Association (2001)
2. Jimmy, Zuccon, G., Koopman, B.: QUT IElab at CLEF 2018 Consumer Health Search Task: Knowledge Base Retrieval for Consumer Health Search. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/paper_203.pdf
3. Mahmood, A.S.M.A., Li, G., Rao, S., McGarvey, P.B., Wu, C.H., Madhavan, S., Vijay-Shanker, K.: UD_GU_BioTM at TREC 2017: Precision Medicine Track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017. vol. Special Publication 500-324. National Institute of Standards and Technology (NIST) (2017), https://trec.nist.gov/pubs/trec26/papers/UD_GU_BioTM-PM.pdf
4. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Hersh, W.R., Bedrick, S., Lazar, A.J.: Overview of the TREC 2018 Precision Medicine Track. In: Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (2018), https://trec.nist.gov/
5. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
6. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 603–610. SIGIR '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1390334.1390437, http://doi.acm.org/10.1145/1390334.1390437

topic	median	TERM_BASED	NO_PRF	PRF
1	0.652	0.571	0.522	0.213
2	0.772	0.841	0.643	0.284
3	0.694	0.714	0.495	0.156
4	0.258	0.262	0.188	0.168
5	0.808	0.765	0.741	0.254
6	0.665	0.688	0.522	0.291
7	0.742	0.736	0.615	0.258
8	0.494	0.602	0.499	0.047
9	0.140	0.175	0.189	0.100
10	0.714	0.833	0.183	0.066
11	0.775	0.828	0.171	0.097
12	0.884	0.922	0.109	0.086
13	0.754	0.894	0.141	0.085
14	0.667	0.670	0.190	0.085
15	0.052	0.031	0.074	0.046
16	0.000	0.000	0.000	0.000
17	0.275	0.353	0.062	0.000
18	0.260	0.568	0.445	0.122
19	0.264	0.521	0.379	0.155
20	0.034	0.305	0.039	0.000
21	0.429	0.648	0.429	0.049
22	0.302	0.451	0.438	0.115
23	0.417	0.459	0.104	0.016
24	0.000	0.000	0.000	0.000
25	0.071	0.056	0.000	0.000
26	0.711	0.801	0.519	0.242
27	0.628	0.722	0.598	0.494
28	0.534	0.487	0.313	0.318
29	0.315	0.669	0.340	0.086
30	0.776	0.861	0.386	0.134
31	0.075	0.000	0.201	0.087
32	0.234	0.364	0.350	0.073
33	0.344	0.575	0.202	0.158
34	0.365	0.593	0.078	0.082
35	0.150	0.791	0.092	0.000
36	0.089	0.097	0.026	0.025
37	0.238	0.507	0.112	0.228
38	0.523	0.514	0.405	0.254
39	0.549	0.636	0.431	0.345
40	0.621	0.689	0.442	0.050
41	0.389	0.569	0.100	0.029
42	0.363	0.446	0.128	0.000
43	0.577	0.577	0.732	0.492
44	0.412	0.846	0.508	0.048
45	0.600	0.637	0.303	0.152
46	0.480	0.585	0.220	0.209
47	0.550	0.798	0.410	0.170
48	0.473	0.646	0.570	0.597
49	0.000	0.150	0.153	0.126
50	0.368	0.525	0.231	0.254
average	0.430	0.540	0.301	0.147

Table 2: Topic by topic comparison of infNDCG values.

topic	P@10	TERM_BASED	NO_PRF	PRF
1	0.800	0.700	0.700	0.100
2	1.000	1.000	1.000	0.300
3	0.900	0.800	0.700	0.200
4	0.300	0.300	0.100	0.400
5	0.900	1.000	0.800	0.100
6	0.800	0.900	0.500	0.400
7	0.700	0.300	0.700	0.200
8	0.600	0.700	0.500	0.000
9	0.300	0.300	0.100	0.100
10	0.900	1.000	0.000	0.000
11	0.800	0.800	0.200	0.000
12	0.900	1.000	0.000	0.000
13	0.900	1.000	0.000	0.000
14	0.800	0.900	0.200	0.000
15	0.000	0.000	0.000	0.000
16	0.000	0.000	0.000	0.000
17	0.300	0.500	0.000	0.000
18	0.200	0.600	0.400	0.100
19	0.200	0.300	0.300	0.100
20	0.000	0.200	0.000	0.000
21	0.700	0.900	0.700	0.000
22	0.400	0.800	0.200	0.000
23	0.500	0.600	0.100	0.000
24	0.000	0.000	0.000	0.000
25	0.000	0.000	0.000	0.000
26	0.800	1.000	0.700	0.400
27	0.700	0.800	0.700	0.700
28	0.400	0.500	0.300	0.200
29	0.700	0.900	0.300	0.100
30	0.700	0.700	0.400	0.000
31	0.000	0.000	0.000	0.100
32	0.300	0.800	0.600	0.000
33	0.600	1.000	0.400	0.200
34	0.100	0.100	0.000	0.000
35	0.100	0.200	0.000	0.000
36	0.100	0.200	0.000	0.000
37	0.400	0.700	0.100	0.400
38	0.500	0.600	0.200	0.100
39	0.400	0.600	0.200	0.100
40	0.900	1.000	0.700	0.200
41	0.200	0.300	0.000	0.000
42	0.100	0.100	0.000	0.000
43	0.500	0.600	0.400	0.400
44	0.500	0.600	0.600	0.200
45	0.600	0.600	0.200	0.100
46	0.600	0.700	0.300	0.100
47	0.600	0.900	0.700	0.100
48	0.300	0.400	0.500	0.500
49	0.000	0.000	0.000	0.000
50	0.400	0.400	0.100	0.000
average	0.468	0.566	0.292	0.118

Table 3: Topic by topic comparison of P@10 values.

topic	median	TERM_BASED	NO_PRF	PRF
1	0.545	0.500	0.436	0.173
2	0.587	0.595	0.500	0.159
3	0.610	0.642	0.423	0.122
4	0.213	0.277	0.170	0.149
5	0.575	0.583	0.475	0.158
6	0.504	0.462	0.403	0.176
7	0.542	0.619	0.381	0.127
8	0.444	0.333	0.296	0.037
9	0.250	0.250	0.125	0.062
10	0.680	0.720	0.040	0.000
11	0.636	0.727	0.136	0.000
12	0.708	0.750	0.000	0.000
13	0.667	0.778	0.037	0.000
14	0.607	0.643	0.143	0.000
15	0.000	0.000	0.000	0.000
16	0.000	0.000	0.000	0.000
17	0.231	0.269	0.000	0.000
18	0.151	0.364	0.364	0.030
19	0.156	0.281	0.312	0.062
20	0.000	0.250	0.000	0.000
21	0.409	0.576	0.424	0.030
22	0.279	0.512	0.349	0.000
23	0.346	0.346	0.077	0.000
24	0.000	0.000	0.000	0.000
25	0.000	0.000	0.000	0.000
26	0.568	0.703	0.324	0.162
27	0.480	0.560	0.480	0.440
28	0.444	0.556	0.222	0.222
29	0.231	0.483	0.187	0.088
30	0.636	0.773	0.182	0.136
31	0.000	0.000	0.000	0.091
32	0.164	0.382	0.309	0.018
33	0.270	0.460	0.149	0.135
34	0.200	0.200	0.000	0.000
35	0.000	0.500	0.000	0.000
36	0.081	0.108	0.027	0.027
37	0.169	0.277	0.139	0.169
38	0.455	0.545	0.182	0.091
39	0.235	0.412	0.176	0.118
40	0.350	0.496	0.219	0.044
41	0.143	0.214	0.000	0.000
42	0.200	0.200	0.000	0.000
43	0.500	0.600	0.400	0.400
44	0.236	0.538	0.274	0.038
45	0.391	0.391	0.174	0.043
46	0.444	0.500	0.167	0.167
47	0.447	0.617	0.340	0.128
48	0.250	0.250	0.312	0.438
49	0.000	0.000	0.000	0.000
50	0.300	0.400	0.200	0.050
average	0.327	0.413	0.191	0.086

Table 4: Topic by topic comparison of R-precision values.