POZNAN Contribution to TREC PM 2018 (Notebook Paper)

Jakub Dutkiewicz[2], Czesław Jędrzejek[2] Artur Cieślewicz[1]

[1] Department of Clinical Pharmacology, Poznan University of Medical Sciences, Poznan, Poland

[2] IARiII, Poznan University of Technology, Poznan, Poland

artcies@ump.edu.pl; {jakub.dutkiewicz, czeslaw.jedrzejek}@put.poznan.pl

Abstract. This work describes the medical information retrieval systems designed by the Poznan Consortium for TREC PM, personalized medicine track, which was submitted to the TREC 2018. The baseline is the Terrier BB2. For Clinical Trials this work uses the following options:

- Terrier BB2_simple_noprf: query without extension (sq_nprf)
- Terrier BB2_simple_w2v_prf: query extended with w2v and terrier prf (sqw2vprf)
- Terrier BB2 _variant_noprf: query + gene variant; without extension (vq_noprf)
- BB2_variant_w2v_prf: query + gene variant extended with w2v and terrier prf (vqw2vprf)
- SQ_results: experimental search of the sql database using keywords from query (sq)

Our best result is vq_noprf which is significantly better (approximately 0.08 for evaluated measures). With suitable word2method the results are better compared to query extension using Mesh and disease taxonomy. For Medline abstracts the documents are placed in the SQL database. For each document templates of diseases, genes, and applicability as Precision Medicine vs research objects are matched. Patterns are saved using regular expressions. The pattern association with the document is stored in the SQL database. For Medline abstracts our results are little worse than the TREC 2018 median.

1 Introduction

The TREC PM 2018 [TREC PM. 2018], is following three previous Clinical Decision Support Track contests and in particular, the TREC PM 2017 [Roberts et al., 2017] . Its aim was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. The 2018 track focuses on an important use case in clinical decision support: providing useful precision medicine-related information to clinicians treating cancer patients. The 2018 track largely repeats the structure and evaluation of the PM 2017 track. TREC 2018 Precision Medicine Track developed Relevance Judgment Guidelines [Relevance Judgment, 2017], the more specific rules compared to TREC CDS to qualify an answer as definitely relevant.

1.1 Source and target documents and rules

There are two target document collections for the Precision Medicine track: scientific abstracts (a January 2017 snapshot of PubMed abstracts plus from AACR and ASCO proceedings targeted toward cancer therapy) and clinical trials (an April 2017 snapshot of ClinicalTrials.gov). Although we submitted runs in both areas, our main concentration was clinical trials task. Topics' description for clinical trials was structured information. Each topic has four primary fields: *Disease*, *Gene* (*Variant)*, and *Demographic*. For instance:

```
<topics task="2018 TREC Precision Medicine">
<topic number="1">
<disease>melanoma</disease>
<gene>BRAF (V600E)</gene>
<demographic>64-year-old male</demographic>
</topic>
```

This structured information allows for a precise definition of „Definitely Relevant" answer: a result should have *Disease* in {*Exact*, *More General*, *More Specific* categories}, at least one *Gene* is *Exact*, and both *Demographic* and *Other* are in. For some topics the gene field contains a specification of the gene action or indirect gene function (Table 1).

Table 1 Additional/alternative gene functions.

Gene function	Topic number
amplification	7, 11, 14
loss of function	5, 17, 23, 24
truncation	15
(extensive) tumor infiltrating lymphocytes	21, 22
high tumor mutational burden	20
rearrangement	16
tumor cells with >50% membranous PD-L1 expression	18
tumor cells negative for PD-L1 expression	19
high serum LDH levels	25

The "(extensive) tumor infiltrating lymphocytes" description can be only attributed to genes, though it is not direct. E.g. the correlation between high TIL levels and improved clinical outcome was most prominent in triple-negative breast cancer (TNBC).

```
<topic number="18">
<disease>melanoma</disease>
<gene>tumor cells with >50% membranous PD-L1 expression</gene>
<demographic>48-year-old female</demographic>
</topic>

<topic number="16">
<disease>melanoma</disease>
<gene>NTRK1 rearrangement</gene>
<demographic>60-year-old male</demographic>
</topic>
<topic number="20">
<disease>melanoma</disease>
<gene>high tumor mutational burden</gene>
<demographic>86-year-old female</demographic>
</topic>
<topic number="24">
<disease>melanoma</disease>
<gene>APC loss of function</gene>
<demographic>47-year-old male</demographic>
</topic>
<topic number="31">
<disease>head and neck squamous cell carcinoma</disease>
<gene>CDKN2A</gene>
<demographic>64-year-old male</demographic>
</topic>
```

Topics 1-6 and 8-13 were characterized by a gene variant. The average of the best TREC PM results as provided by organizers for them is 0.80 compared with the same result for the rest of topics which is 0.70.

1.2 Experience

The Poznan Consortium team for TREC PM consists of contributors of two institutions: Department of Clinical Pharmacology, Poznan University of Medical Sciences, and Information Systems and Technologies Division, IARiII, Poznan University of

Technology. We participated in TREC CDS 2016 track [Dutkiewicz et al, 2017], TREC PM 2017 track [Cieslewicz et al. , 2018a] and in bioCADDIE 2016 [Cieslewicz et al. , 2018b]. In this work we use two type of embedding: designed for biomedical applications [Chiu et al., 2016] and classical word embedding [Mikolov et al., 2013a] on the Pubmed abstracts corpus. We did not preprocess the target Open Access Subset of PubMed Central (PMC).

2 Related work

2.1 Baseline system

Our experience shows that for biomedical systems Terrier 4.2 is among the best for baseline systems. The best performing depending were 1) BB2 (Bernoulli-Einstein model with Bernoulli after-effect and normalization [Amati, van Rijsbergen, 2012], also denoted as "DPH + Bo1) [Dutkiewicz, 2017], 2) LGD [Clinchant, Gaussier, 2010] (a log-logistic model for information retrieval) [Cieslewicz, 2018b]. Another valuable feature implemented in Terrier is pseudo relevance feedback query expansion (PRF) – a mechanism allowing for extraction of *n* most informative terms from *m* top ranked documents (ranking created in the first search run) which are then added to the original query in the second retrieval rank. Terrier provides both parameter free (Bose-Einstein 1, Bose-Einstein 2, Kullback-Leibler) and parameterized (Rocchio) models for query expansion [Rocchio, 1971]. Rocchio feedback approach was developed using the Vector Space Model. The modified vectors are moved in a direction closer or farther away, from the original query depending on whether documents are related or non-related.

2.2 Query expansion

Expanding queries by adding potentially relevant terms is a common practice in improving relevance in IR systems. There are many methods of query expansion. Relevance feedback takes the documents on top of a ranking list and adds terms appearing in these document to a new query. In this work we use the idea to add synonyms and other similar terms to query terms before the pseudo-relevance feedback. This type of expansion can be divided into two categories. The first category involves the use of ontologies or lexicons (relational knowledge). In biomedical area UMLS, MeSH [MeSH, 2018], SNOMED-CT, ICD-10, WordNet, and Wikipedia are used [JaiswaL, 2016]. Generally, the result of lexicon type of expansion is positive. The second category is word embedding (WE). This belongs to a class of distributional semantics, feature learning techniques in natural language processing. One can draw experience on effect of using lexicons from other semantic task areas.

For natural language queries requiring an answer using multiple choice, relational learning using dictionaries encompassing the whole corpus gives always better results than pure word embedding (word2vec). However, having synonym dictionaries (flat structure) can significantly improve the word2vec results.

3 Retrieval methodology setup

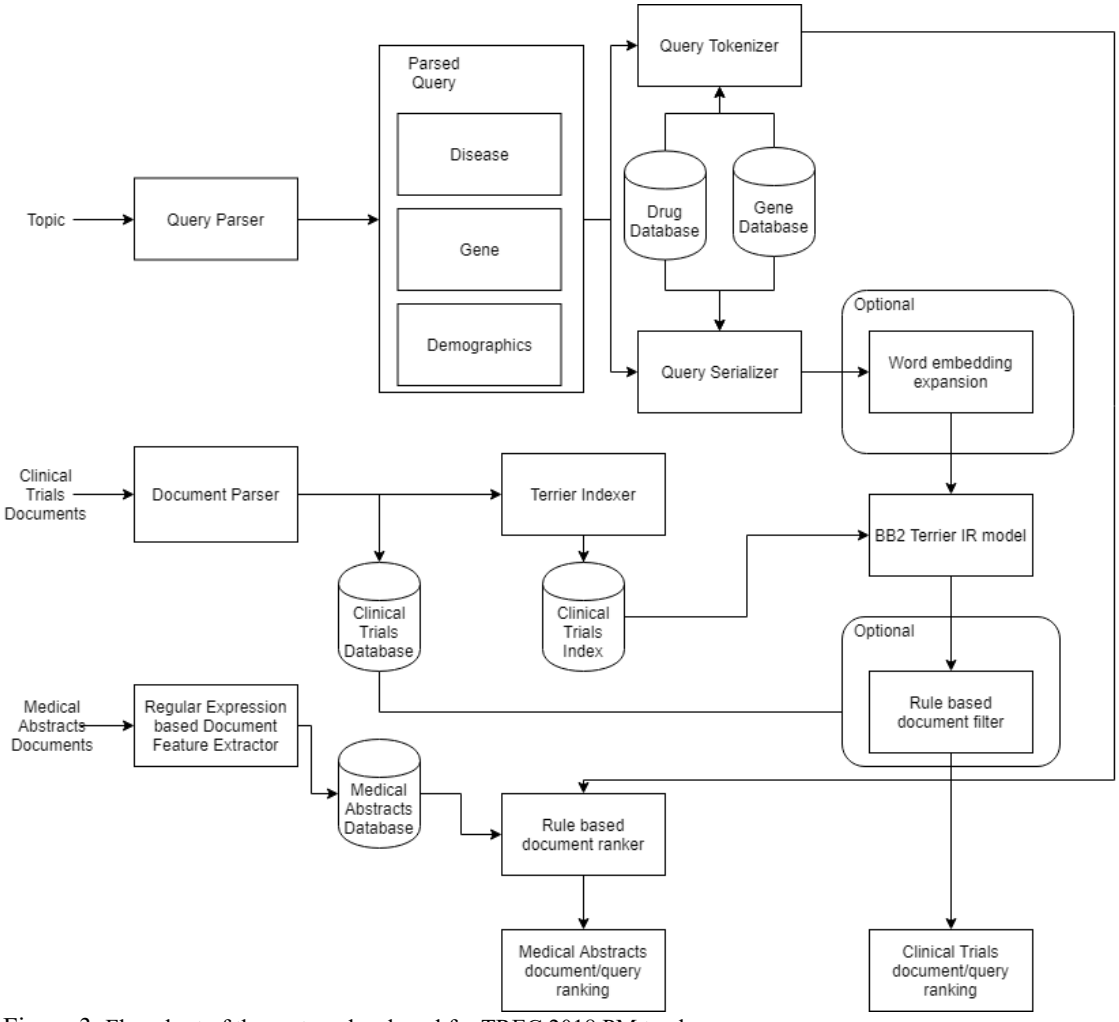

Figure 3. Flowchart of the system developed for TREC 2018 PM track.

3.1 Methodology developed for Medline abstracts

1. Data indexing.

   For the set of medical abstract documents, we perform regular expression based feature extraction. We look for documents which relate to various genes, diseases, drugs which may relate to diseases within the query set or gene variants, by applying the dedicated pattern to the document. Regex patterns relate to the features and contain a variety of syntactic irregularities, such as:

   - -delimiter between the main part of the gene/gene variant name and a number,
   - -various letter case options,
   - various gene/gene variant/disease/drug synonyms.

   Exact value of a feature is calculated as number of the regular expression matches divided by the document length (in words).

2. Query construction.

   Queries are translated into SQL SELECT statements. The generated statements are:

   a. Select documents which are classified as related to a given disease
   b. Select documents which contain information about a given gene

c. Select documents which contain information about a given variant
d. Select a specific value of a feature for a given document

3. Information retrieval

We perform a single run "boolean" on the set. A document is taken into relevant document list only if it contains at least one gene name named in the query and it contains one of the disease name synonyms. We assume that a document, which relates to a disease is a Precision Medicine document by default. We rank these documents which contain correct gene variant higher than documents which miss the gene variant and higher than documents with incorrect gene variant. Further ranking is calculated upon the values of the related features - the higher the combined sum of features is, the higher ranking score the document receives.

3.2 Methodology developed for Clinical Trials data

1. Data indexing.

The content of clinical trials documents was indexed as follows:

- tag DOCNO: the value of nct_id tag
- tag <TEXT>: the values of tags brief_title, official_title, brief_summary, detailed_description, primary_outcome, secondary_outcome, condition, arm_group, condition_browse, intervention_browse, keyword; additionally, inclusion criteria were extracted from criteria tag

2. Query construction.

Three types of queries were constructed:

1. simple query: find documents with any term from <disease> tag or any gene name from <gene> tag; gene variants were not taken into consideration
2. variant query: find documents with any term from <disease> tag or any term from <gene> tag; gene variants were taken into consideration
3. SQL query: browse data from SQL database to find records that have any term from <disease> tag and any term from <gene> tag; gene variants were not taken into consideration

Queries 1) and 2 were also expanded in a following manner: terms from simple or variant query were expanded using 2 vector space models (word2vec): first calculated by us on the basis of bioCADDIE corpus; second calculated by [Chiu , et al., 2016] on the text corpus based on PubMed citations and full-text articles
As a result, five queries were constructed for each topic.

3. Information retrieval

4 terrier runs were carried out, using BB2 as ranking function:

1. BB2_simple_noprf: simple query was put as input for terrier
2. BB2_simple_w2v_prf: simple query expanded with word2vec was put as input for terrier; additional expansion was carried out using terrier pseudo relevance feedback (PRF)
3. BB2_variant_noprf: variant query was put as input for terrier

4. BB2_variant_w2v_prf: variant query expanded with word2vec was put as input for terrier; additional expansion was carried out using terrier PRF

The last run (SQ_results) was performer as follows:

5. data from <nct_id>, <brief_title>, <official_title>, <brief_summary>, <detailed_description>, <keyword>, <condition_browse>, and <study_type> xml tags of each document was put into SQL database; inclusion criteria were also extracted from <criteria> tag and were placed in the database in "inclusion" field for each topic, an SQL query (described in "Query construction" section) was executed obtain results were then ranked, according to the number of occurrences of disease, gene name and gene variant terms in each record fields; documents describing international studies were granted additional points to total score

All generated results were then checked according to the value of <demographic> tag of each topic. Resulting documents that were describing trials recruiting patients with inappropriate age or gender were removed from the result set.

4 Results

In this section we go through the outcome of every retrieval setup implemented by our group and applied to the competition data sets. We compare our results to median and best of the PM submissions. Finally, we discuss the best application for each setup. For the evaluation we show measures that were used by TREC PM evaluators for abstracts and clinical trials.

4.1 Results for Medline abstracts

In this category we submitted one run using Boolean search. Our average result is little worse than TREC PM median.

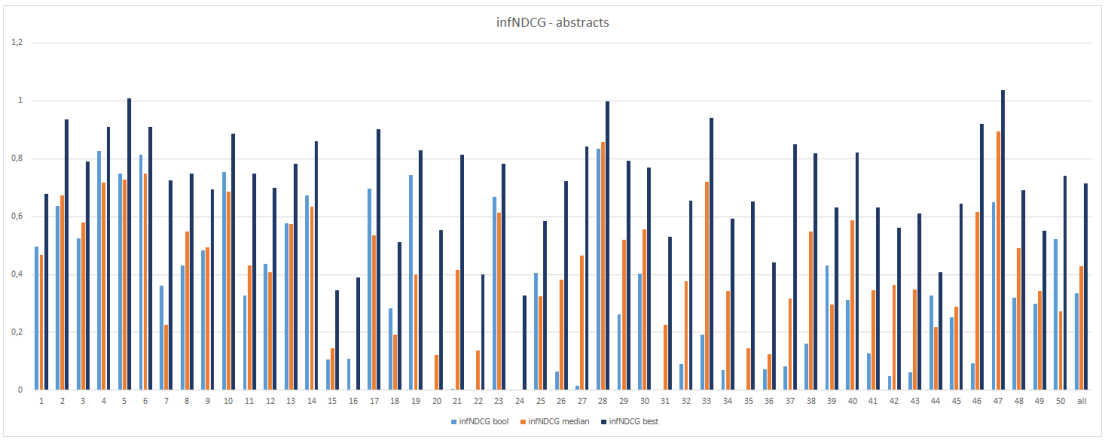

Figure 4.1 infNDCG results for Medline abstracts for individual topics and median averages (ours and the TREC PM 2018).

4.2 Results for Clinical Trials data

In this category we submitted 5 runs, all using BB2 as a baseline. In TREC PM 2017 we used the LGD option that was the best for bioCADDIE 2016 challenge. For the purpose of this work, having qrels for the last year TRC PM we verified that BB2 method beats the LGD one.

The results generated for Clinical trials data have shown that using Pseudo Relevance Feedback (PRF) to expand the query had negative impact for almost all measures (see the Table 4.1). Terrier PRF was configured to expand queries with 10 terms from top 2 documents. Observed worsening of the results could be explained with the fact that PRF was carried out before documents describing trials recruiting patients with inappropriate age or gender were removed from the result set. Another aspect worth

considering is that Clinical trials information retrieval required finding the documents describing clinical trials for which the patient described in the query is eligible for the recruitment. PRF, by adding additional terms to the query, could improve the score of not relevant documents.

The best results were obtained with word embedding using both vectors obtained by [Chiu et al., 2016] and by our classical [Mikolov, et al., 2013] type of calculations with vector cosine measure = 0.875.

Green color indicates the results better than 0.5, red color the results worse than 0.5 than median results.

Table 4.1 Results for various options for the provided Clinical Trials data runs (intensity of these colors proportional to divergence from the medians).

topic	sq_nprf	vq_noprf	vqw2vprf	sqw2vprf	sq	trec_best	trec_median
1	0.613	0.806	0.626	0.6239	0.7373	0.8489	0.6516
2	0.7232	0.7913	0.7334	0.7295	0.7724	0.9054	0.7724
3	0.6989	0.7247	0.699	0.7119	0.6665	0.8424	0.6937
4	0.3416	0.3416	0.3416	0.3231	0.2157	0.4584	0.2583
5	0.859	0.9239	0.8797	0.8593	0.8885	0.9645	0.8084
6	0.7031	0.7531	0.774	0.6924	0.7488	0.8685	0.6651
7	0.8326	0.8565	0.8646	0.8323	0.7893	0.9652	0.7421
8	0.7013	0.7013	0.7013	0.5796	0.4353	1.0942	0.4944
9	0.1323	0.1323	0.1323	0.2658	0.1385	0.2658	0.1396
10	0.7195	0.7195	0.7195	0.6899	0.6886	0.9078	0.7135
11	0.8052	0.9154	0.8104	0.7709	0.8272	0.9291	0.7746
12	0.9055	0.9055	0.9055	0.8497	0.935	0.9627	0.8837
13	0.8697	0.8697	0.8697	0.848	0.8141	0.8941	0.7536
14	0.7874	0.7256	0.7266	0.7506	0.6935	0.8067	0.6667
15	0.0895	0.0893	0.0875	0.0411	0	0.2744	0.0518
16	0	0	0	0	0	0.5	0
17	0.2215	0.2617	0.259	0.1949	0.2962	0.608	0.2754
18	0.1249	0.0555	0	0	0.4544	0.6743	0.26
19	0.242	0.2635	0.021	0	0.4082	0.6679	0.2635
20	0	0	0	0	0	0.5027	0.0337
21	0.0193	0.6864	0	0.017	0.6035	0.8145	0.4292
22	0.0116	0.4929	0	0.0108	0.3634	0.5291	0.3016
23	0.4168	0.4363	0.4321	0.4326	0.3356	0.6817	0.417
24	0	0	0	0	0	0.6309	0
25	0.0781	0.0708	0.0708	0.0868	0	0.2934	0.0708
26	0.7522	0.7522	0.7517	0.743	0.7809	0.8929	0.7112
27	0.7732	0.7732	0.7601	0.8193	0.2841	0.9098	0.6276
28	0.4738	0.4738	0.4738	0.4774	0.4579	0.7743	0.5339
29	0.5055	0.5055	0.5021	0.5001	0.3899	0.7313	0.3151
30	0.8085	0.8085	0.8089	0.8373	0.6184	0.904	0.7763
31	0	0	0	0	0	0.469	0.0746
32	0.1631	0.1631	0.1504	0.4484	0.0826	0.9194	0.234
33	0.522	0.522	0.5204	0.5209	0.5029	0.7171	0.3435
34	0.4972	0.4972	0.5633	0.4693	0.4453	0.7263	0.3654
35	0.5798	0.5798	0.5798	0.3468	0	0.842	0.1505
36	0.0896	0.0896	0.0896	0.082	0.1232	0.7636	0.0887
37	0.0188	0.0188	0.0252	0	0.0503	0.6119	0.2384
38	0.5231	0.5231	0.5231	0.5558	0.553	0.6993	0.5231
39	0.6131	0.6131	0.2519	0.2034	0.6262	0.868	0.5486
40	0.7283	0.7283	0.7283	0.7201	0.7705	0.8546	0.6206
41	0.7082	0.7082	0.6903	0.8744	0	0.8744	0.389
42	0.363	0.363	0.363	0.363	0.363	0.7735	0.363
43	0.6545	0.6545	0.6546	0.6223	0.6793	0.8021	0.5773
44	0.0779	0.0779	0.0779	0.0779	0.0922	0.9987	0.4119
45	0.7108	0.7108	0.7095	0.6527	0.5548	1.1734	0.5996
46	0.6196	0.6196	0.6196	0.6139	0.5447	0.7885	0.4801

47	0.727	0.727	0.727	0.7496	0.5802	0.8101	0.55
48	0.4825	0.4825	0.4825	0.571	0.3536	0.6527	0.4727
49	0.1205	0.1205	0.15	0.147	0	0.6131	0
50	0.4342	0.4342	0.4381	0.456	0.5999	0.7341	0.3676
all	0.4568	0.4894	0.4459	0.4432	0.4253	0.755894	0.429668

Figures 4.2-4.4 present results for particular topics compared with the best and median TREC PM 2018 results.

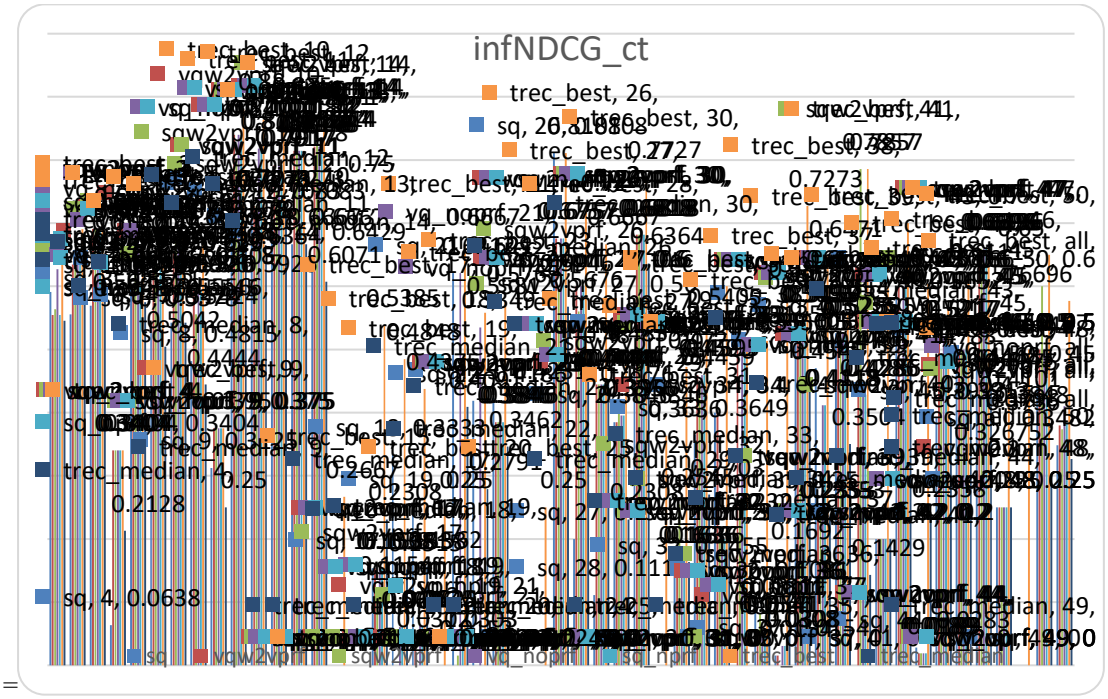

Figure 4.2 infNDCG results for Clinical Trials for individual topics and median averages (ours and the TREC PM 2018).

Figure 4.3 R-prec_ct for Clinical Trials for individual topics and median averages (ours and the TREC PM 2018).

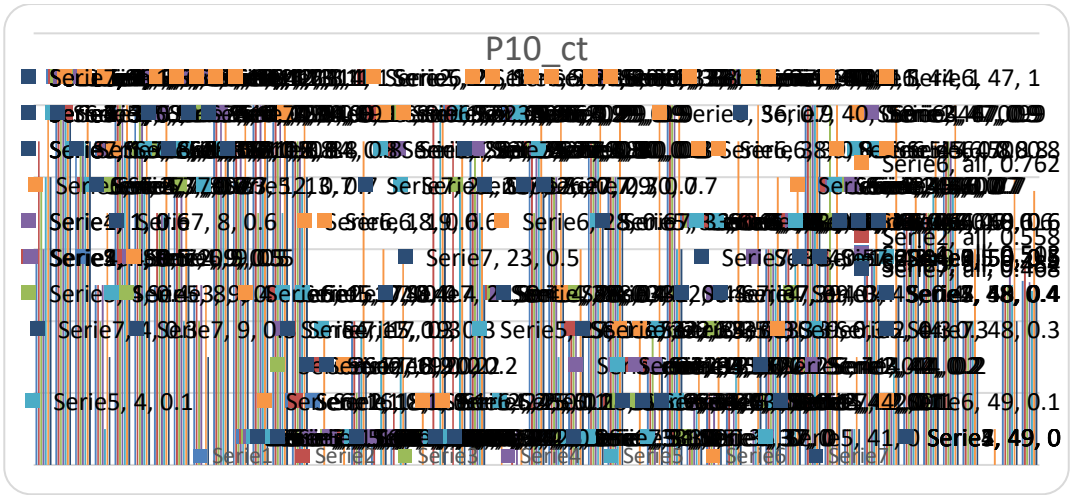

Figure 4.4 P-10_ct for Clinical Trials for individual topics and median averages (ours and the TREC PM 2018).

4 Conclusions and future work

For Medline abstracts our results are little worse than the TREC 2018 median. The method is different from classical methods, and similar to Boolean searches. The biggest deficiency of our approach for Medline abstracts is lack of distinction between scientific papers and personalized treatment papers. Presumably this feature explains very good results of [Goodwin et al., 2017] and [Mahmood et al., 2017] in the TREC PM 2017 track. Another reason could be extraction of relations used be these two teams [Peng et al., 2016]. For Clinical Trials our best result is vq_noprf which is significantly better (approximately 0.08 for evaluated measures). With a suitable word2vec method the results are better compared to query extension using Mesh and disease taxonomies.

References

[Amati, van Rijsbergen, 2012] Amati,G., van Rijsbergen,C. J., (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inf. Syst. 20(4): 357-389

[TREC PM. 2018], [http://www.trec-cds.org/2018.html

[TREC PM, Relevance guidelines, http://www.trec-cds.org/relevance_guidelines.pdf

[Roberts et al, 2017] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees William R. Hersh and Steven Bedrick Alexander J. Lazar Shubham Pant, Overview of the TREC 2017 Precision Medicine Track
https://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf

[Chiu , et al., 2016] Chiu,B., Crichton,G., Korhonen,A. et al. (2016) How to train good word embeddings for biomedical NLP. In: Proceedings of the 5th Workshop on Biomedical Natural Language Processing. Berlin, Germany. http://www.aclweb.org/anthology/W16-2922.

[Cieslewicz et al, 2017a] Artur Cieslewicz, Jakub Dutkiewicz, Czeslaw Jedrzejek: Baseline and extensions approach to information retrieval of complex medical data: Poznan's approach to the bioCADDIE 2016. Database 2018: bax103 (2018).

[Cieslewicz et al, 2017b] Artur Cieslewicz, Jakub Dutkiewicz, Czeslaw Jedrzejek:, POZNAN Contribution to TREC PM 2017, https://trec.nist.gov/pubs/trec26/papers/POZNAN-PM.pdf

[Clinchant, Gaussier,2010] Clinchant,S., Gaussier,E., (2010) Information-based models for ad hoc IR. In *SIGIR'10*, 234-241.

[Dutkiewicz et al, 2017] Jakub Dutkiewicz, Czesław Jędrzejek, Michał Frąckowiak and Paweł Werda, PUT Contribution to TREC CDS 2016, https://trec.nist.gov/pubs/trec25/papers/IAII_PUT-CL.pdf

Terrier 5.0 IR Platform *www.terrier.org http://terrier.org/docs/v5.0/* accessed July 2, 2018

[Goodwin et al., 2017] Travis R. Goodwin, Michael A. Skinner and Sanda M. Harabagiu UTD HLTRI at TREC 2017: Precision Medicine Track, https://trec.nist.gov/pubs/trec26/papers/UTDHLTRI-PM.pdf

[Mahmood et al., 2017] A. S. M. Ashique Mahmood, Gang Li , Shruti Rao , Peter McGarvey, Cathy Wu, Subha Madhavan2, K. Vijay-Shanker UD_GU_BioTM at TREC 2017: Precision Medicine Track
https://trec.nist.gov/pubs/trec26/papers/UD_GU_BioTM-PM.pdf

[Jaiswa et al., 2016] Jaiswal,P., Hoehndorf,R., Cecilia,N. et al. (2016) Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States. CEUR Workshop
Proceedings 1747, CEUR-WS.org 201.

[Mikolov, et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, Corrado, G.S., and Dean, (2013b), J. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781

[MeSH, 2018], MeSH database: https://www.nlm.nih.gov/mesh/download_mesh.html, Access: July 2, 2018

[Peng et al., 2016] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih:
Cross-Sentence N-ary Relation Extraction with Graph LSTMs. TACL 5: 101-115 (2017)

[Rocchio, 1971] Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system—Experiments in automatic document processing* (pp. 313–323). New York City, USA: Prentice-Hall.