POZNAN Contribution to TREC PM 2017

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

[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 2017. The baseline is the Terrier DPH Bo1 which recently has been shown to be the most effective Terrier option. We also used Mesh query expansion, word2vec query expansion, and the combination of these two options. In all measures our results are approximately 0,02 above the median.

1. Introduction

The TREC PM 2017 is following three previous Clinical Decision Support Track contests. Its aim was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. This year's track was dedicated to the personalization aspect of retrieved information, namely, „ *if an abstract provides information relevant to the treatment of the patient's cancer?"* or *"is a patient eligible for this clinical trial*?".

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 contributions in both area our main concentration was clinical trials. Topics' description for clinical trials was structured information. Each topic has four primary fields: *Disease*, *Variant*, *Demographic*, and *Other*. For instance:

Disease: Acute lymphoblastic leukemia **Variant**: ABL1, PTPN11 **Demographic**: 12-year-old male **Other**: None. The „**Other" could contain additional diseases of a patient or ECOG p**erformance status.

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.

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 and in bioCADDIE 2016 [Cieslewicz, 2017]. Dutkiewicz, Jedrzejek, 2017]. In our research we use the word embedding [Mikolov et al., 2013a], with semantic (relation) knowledge [Faruqui et.al., 2015], [Dutkiewicz, Jedrzejek, 2018]. We did not preprocess the target Open Access Subset of PubMed Central (PMC).

2. Baseline System

Our experience shows that for biomedical systems Terrier 4.1 is among the best for baseline systems. The best performing were 1) BB2 (Bernoulli-Einstein model with Bernoulli after-effect and normalization, also denoted as "DPH + Bo1 2), LGD (a log-logistic model for information retrieval) [Cieslewicz, 2017]. 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 (24). 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.

2.1. 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 (20), SNOMED-CT, ICD-10, WordNet, and Wikipedia are used (27). 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.

In this contribution we do not use word embedding for query expansion.

3. Retrieval methodology setup

We are using a dedicated retrieval methodology. We process and index metadata provided by Clinical Trials articles. We use manually constructed gene knowledge base to expand the query. We index all articles from Clinical Trials and Scientific Medline Abstracts into two separate data entities. Algorithms are described in the following sections.

3.1. Methodology developed for Medline Abstracts

The Medline documents were converted into a dedicated XML format as follows:

1. Value of <DOCNO> : the value of <PMID> tag,
2. Value of <TEXT> : concatenation of vales of tags <KEYWORDS> <ARTICLETITLE> <ABSTRACT> and <MESHHEADING>.

Queries were constructed with two different setups:

1. simple: find documents with any term from <disease>, <gene> or <other> tag present in <TEXT> field of indexed content,
2. strict: find documents with all terms from <disease> and <gene> tag and any term from <other> tag present in <TEXT> field of indexed content.

"strict" queries were expanded manually in the same way as described in the method developed for Clinical Trials.

Medline documents were indexed with Terrier and 3 retrieval runs were carried out:

a. POZabsBB2sn: simple query was used as input; ranking function was BB2
b. POZabsBB2GRn: a set of "strict" queries (generated with manual expansion) was used as input; ranking function was BB2
c. POZabsLGDGRn: a set of "strict" queries (generated with manual expansion) was used as input; ranking function was LGD

For runs b and c, generated results contained duplicates, which were removed by taking only one document with highest score.

3.2 Methodology developed for Clinical Trials data

The Medline documents were converted into a dedicated XML format as follows:

1. tag <DOCNO>: the value of <*nct_id*> tag
2. 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
3. tag <NEGATIVE>: exclusion criteria extracted from <*criteria*> tag

Three types of queries were constructed, based on the topic structure:

a. simple: find documents with any term from <disease> or <gene> tag present in <TEXT> field of indexed content and no term from <other> tag present in <NEGATIVE> field of indexed content,
b. strict: find documents with all terms from <disease> and <gene> tag present in <TEXT> field of indexed content and no term from <other> tag present in <NEGATIVE> field of indexed content,
c. optional gene: find documents with all terms from <disease> tag present in <TEXT> field of indexed content, any term from <gene> tag present in <TEXT> field of indexed content, and no term from <other> tag present in <NEGATIVE> field of indexed content.

b) and c) queries were manually expanded in a following manner:

b. for each gene name a list of synonymous gene names was prepared (based on the data from NCBI Gene database: ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz,
c. for each disease term a list of synonymous names was prepared (based on entry terms from NCBI MeSH database)

For each query a set of queries was produced based on prepared synonymous terms.
Five Terrier runs were carried out, using LGD as ranking function:

a. LGDraw: “simple” query was put as input for terrier; the query was expanded with terrier pseudo relevance feedback (PRF)
b. LGDStrict: a set of “strict” queries (generated with manual expansion) was used as input
c. LDGprfStrict: as LGDStrict but with terrier PRF
d. LGDnoprfGOpt: a set of “optional gene” queries (generated with manual expansion) was used as input
e. LGDprfGOpt: as LGDnoprfGOpt but with terrier PRF
f. For runs b-e, generated results contained duplicates, which were removed by taking only one document with highest 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 3 runs. Summary for the runs we provided is presented in Table 4.1. BB2 with no query expansion is little better than TREC PM median and much better than BB2 with query expansion and LGD with query expansion.

4.2 Results for Clinical Trials

In this category we submitted 5 runs, all using LGD as a baseline. Summary for the runs we provided is presented in Table 4.2. Similarly to the Medline Abstracts results, baseline LGD performs better that any expansion option and is much better than TREC PM median. The gain over TREC PM median is 27% for P5 measure, 47% for P10 measure, and 49% for P15 measure.

The results generated for Clinical trials data have shown that using Pseudo Relevance Feedback (PRF) to expand the query had negative impact on almost all measures (see the Table 4.3). 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 therefore improve the score of not relevant documents.

Table 4.3 The effect of PRF on P5, P10 and P15 measure.

Measure	LGD strict	LGD strict + PRF	Difference(%)	LGD opt	LGD opt + PRF	Difference(%)
P5	0,386	0,293	-24,07%	0,379	0,393	+3,64%
P10	0,282	0,257	-8,85%	0,336	0,311	-8,05%
P15	0,267	0,221	-16,98%	0,295	0,262	-12,71%

Table 4.1 Summary for the provided Medline abstracts runs.

TOPIC / MEASURE	METHOD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	ALL TOPICS
infNDCG	best TREC	0,61	0,88	0,51	0,83	0,30	0,57	0,70	0,56	0,85	0,39	0,79	0,00	0,32	0,67	0,51	0,69	0,53	0,50	0,47	0,48	0,60	0,64	0,91	0,56	0,53	0,63	0,31	0,48	0,46	0,48	0,56
	median TREC	0,46	0,60	0,28	0,41	0,16	0,42	0,34	0,27	0,64	0,17	0,21	0,00	0,06	0,03	0,13	0,41	0,26	0,31	0,19	0,17	0,38	0,18	0,51	0,25	0,26	0,09	0,10	0,12	0,21	0,19	0,26
	absLGDGRn	0,40	0,50	0,10	0,50	0,12	0,22	0,13	0,23	0,55	0,14	0,31	0,00	0,00	0,11	0,25	0,32	0,30	0,35	0,21	0,18	0,36	0,30	0,17	0,05	0,03	0,00	0,09	0,11	0,18	0,05	0,21
	absBB2sn	0,41	0,55	0,28	0,78	0,20	0,38	0,22	0,32	0,66	0,20	0,31	0,00	0,00	0,01	0,36	0,40	0,27	0,46	0,23	0,17	0,49	0,36	0,74	0,08	0,00	0,12	0,09	0,04	0,11	0,29	0,28
	absBB2GRn	0,40	0,47	0,13	0,58	0,15	0,15	0,16	0,17	0,41	0,17	0,26	0,00	0,00	0,12	0,24	0,45	0,26	0,34	0,22	0,18	0,42	0,38	0,45	0,02	0,04	0,00	0,05	0,13	0,18	0,05	0,22
P10	best TREC	1,00	1,00	0,90	1,00	0,50	0,90	1,00	0,90	1,00	0,60	0,90	0,00	0,50	0,80	0,30	1,00	1,00	1,00	0,90	0,70	0,90	0,90	1,00	1,00	0,70	0,80	0,80	0,90	1,00	0,90	0,83
	median TREC	0,60	0,90	0,30	0,70	0,20	0,60	0,50	0,30	0,80	0,20	0,20	0,00	0,10	0,00	0,10	0,60	0,30	0,50	0,20	0,20	0,50	0,40	0,60	0,40	0,30	0,10	0,10	0,20	0,30	0,20	0,35
	absLGDGRn	0,50	1,00	0,40	0,60	0,20	0,10	0,10	0,50	0,70	0,30	0,70	0,00	0,00	0,20	0,20	0,40	0,50	0,70	0,20	0,30	0,80	0,10	0,00	0,20	0,00	0,00	0,10	0,10	0,30	0,10	0,31
	absBB2sn	0,50	0,80	0,30	0,70	0,30	0,60	0,10	0,40	0,90	0,30	0,90	0,00	0,00	0,00	0,10	1,00	0,10	0,80	0,30	0,10	0,60	0,50	0,90	0,10	0,00	0,10	0,20	0,10	0,20	0,50	0,38
	absBB2GRn	0,40	1,00	0,40	0,70	0,00	0,00	0,20	0,40	0,50	0,30	0,60	0,00	0,00	0,20	0,10	0,70	0,40	0,80	0,30	0,10	0,50	0,70	0,60	0,00	0,00	0,00	0,10	0,10	0,30	0,10	0,32
R-prec	best TREC	0,50	0,44	0,41	0,42	0,21	0,42	0,38	0,34	0,46	0,27	0,60	0,00	0,24	0,55	0,30	0,46	0,37	0,30	0,34	0,35	0,39	0,42	0,57	0,44	0,37	0,50	0,21	0,40	0,40	0,33	0,38
	median TREC	0,34	0,27	0,20	0,21	0,12	0,30	0,17	0,23	0,35	0,13	0,14	0,00	0,04	0,00	0,10	0,26	0,23	0,15	0,09	0,12	0,30	0,11	0,31	0,17	0,17	0,05	0,09	0,07	0,14	0,15	0,17
	absLGDGRn	0,39	0,17	0,08	0,16	0,09	0,21	0,07	0,14	0,19	0,10	0,29	0,00	0,00	0,06	0,20	0,26	0,20	0,10	0,14	0,20	0,24	0,25	0,19	0,03	0,02	0,00	0,05	0,07	0,07	0,03	0,13
	absBB2sn	0,35	0,22	0,20	0,26	0,18	0,26	0,12	0,23	0,31	0,16	0,26	0,00	0,00	0,00	0,10	0,23	0,28	0,19	0,11	0,18	0,27	0,16	0,38	0,03	0,00	0,10	0,07	0,02	0,05	0,19	0,16
	absBB2GRn	0,37	0,16	0,08	0,12	0,10	0,20	0,09	0,12	0,16	0,13	0,21	0,00	0,00	0,06	0,10	0,25	0,12	0,10	0,14	0,18	0,25	0,20	0,17	0,03	0,02	0,00	0,04	0,09	0,07	0,03	0,12

Table 4.2 Summary for the provided Clinical Trials runs.

TOPIC / METHOD	MEASURE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	ALL TOPICS
TREC BEST RESULTS	P5	1,00	1,00	1,00	1,00	0,80	0,80	1,00	0,80	1,00	0,00	0,80	0,00	0,60	1,00	0,40	0,40	0,60	0,60	0,80	0,40	1,00	1,00	0,80	1,00	1,00	0,20	0,80	0,00	0,80	1,00	0,77
	P10	0,90	0,90	1,00	0,90	0,50	0,80	1,00	0,70	1,00	0,00	0,70	0,00	0,60	0,60	0,20	0,30	0,60	0,40	0,50	0,30	1,00	1,00	0,70	1,00	1,00	0,30	0,60	0,10	0,60	0,70	0,68
	P15	0,67	0,93	0,87	0,80	0,53	0,67	1,00	0,67	1,00	0,00	0,67	0,00	0,67	0,40	0,13	0,27	0,53	0,33	0,40	0,27	0,93	0,93	0,53	0,87	0,80	0,20	0,47	0,07	0,40	0,53	0,59
TREC MEDIAN RESULTS	P5	0,80	0,60	0,60	0,40	0,20	0,60	0,60	0,20	0,60	0,00	0,40	0,00	0,00	0,40	0,00	0,20	0,20	0,00	0,20	0,00	0,20	0,40	0,20	0,60	0,40	0,00	0,00	0,00	0,20	0,20	0,29
	P10	0,40	0,60	0,60	0,40	0,20	0,50	0,60	0,20	0,70	0,00	0,20	0,00	0,00	0,30	0,00	0,10	0,20	0,10	0,10	0,00	0,30	0,30	0,10	0,50	0,30	0,00	0,10	0,00	0,10	0,20	0,25
	P15	0,27	0,60	0,47	0,33	0,20	0,40	0,67	0,27	0,67	0,00	0,20	0,00	0,00	0,20	0,00	0,07	0,20	0,07	0,13	0,07	0,27	0,20	0,13	0,33	0,27	0,00	0,13	0,00	0,07	0,13	0,23
LGDprf Strict	P5	0,80	0,20	0,40	0,40	0,20	0,60	0,40	0,20	1,00	0,00	0,20	0,00	0,00	0,60	0,00	0,00	0,00	0,40	0,00	0,20	0,20	0,40	0,00	1,00	0,60	0,00	0,00	0,00	0,20	0,20	0,29
	P10	0,50	0,40	0,20	0,40	0,20	0,50	0,30	0,30	1,00	0,00	0,10	0,00	0,00	0,60	0,00	0,10	0,00	0,20	0,20	0,10	0,20	0,70	0,00	0,60	0,40	0,00	0,00	0,00	0,10	0,10	0,26
	P15	0,33	0,47	0,13	0,33	0,13	0,53	0,27	0,27	0,93	0,00	0,20	0,00	0,00	0,40	0,00	0,20	0,00	0,13	0,13	0,07	0,13	0,60	0,00	0,40	0,27	0,00	0,13	0,00	0,07	0,07	0,22
LGDnoprf Strict	P5	0,80	0,80	0,40	0,80	0,40	0,60	0,20	0,00	1,00	0,00	0,80	0,00	0,00	0,80	0,00	0,40	0,20	0,40	0,40	0,20	0,20	0,60	0,00	0,80	0,60	0,00	0,00	0,00	0,20	0,20	0,39
	P10	0,50	0,90	0,20	0,60	0,20	0,50	0,10	0,10	0,80	0,00	0,40	0,00	0,00	0,60	0,00	0,20	0,10	0,20	0,20	0,10	0,10	0,70	0,10	0,60	0,30	0,10	0,10	0,00	0,10	0,10	0,28
	P15	0,33	0,80	0,13	0,60	0,13	0,53	0,20	0,33	0,73	0,00	0,27	0,00	0,00	0,40	0,00	0,20	0,20	0,13	0,13	0,07	0,20	0,67	0,20	0,47	0,33	0,07	0,20	0,00	0,07	0,07	0,27
LGD	P5	0,60	0,80	0,80	0,40	0,20	0,60	0,40	0,00	0,60	0,00	0,60	0,00	0,00	0,80	0,20	0,20	0,20	0,40	0,40	0,20	0,20	0,60	0,00	1,00	0,40	0,00	0,20	0,00	0,20	0,40	0,37
	P10	0,60	0,70	0,90	0,50	0,20	0,50	0,30	0,20	0,70	0,00	0,40	0,00	0,00	0,50	0,20	0,20	0,40	0,30	0,40	0,10	0,20	0,60	0,10	1,00	0,70	0,00	0,40	0,00	0,10	0,20	0,37
	P15	0,40	0,80	0,80	0,53	0,20	0,53	0,40	0,33	0,80	0,00	0,40	0,00	0,00	0,33	0,13	0,20	0,40	0,20	0,33	0,07	0,20	0,60	0,07	0,67	0,53	0,00	0,33	0,00	0,07	0,13	0,34
LGDprf Opt	P5	0,60	1,00	0,80	0,20	0,40	0,60	0,40	0,20	1,00	0,00	0,20	0,00	0,00	0,60	0,20	0,20	0,00	0,40	0,40	0,20	0,40	0,60	0,00	1,00	0,60	0,00	0,20	0,00	0,20	0,60	0,39
	P10	0,60	0,70	0,80	0,20	0,30	0,50	0,30	0,20	1,00	0,00	0,30	0,00	0,00	0,50	0,10	0,20	0,00	0,20	0,20	0,10	0,20	0,70	0,00	0,60	0,40	0,00	0,10	0,00	0,10	0,40	0,31
	P15	0,40	0,67	0,73	0,27	0,27	0,53	0,27	0,13	0,93	0,00	0,20	0,00	0,00	0,40	0,07	0,20	0,00	0,13	0,13	0,07	0,13	0,53	0,07	0,40	0,27	0,00	0,13	0,00	0,07	0,33	0,26
LGDnoprf Opt	P5	0,60	0,80	0,80	0,20	0,20	0,60	0,20	0,00	1,00	0,00	0,60	0,00	0,00	0,80	0,00	0,40	0,20	0,40	0,40	0,20	0,20	0,60	0,00	0,80	0,60	0,00	0,00	0,00	0,20	0,80	0,38
	P10	0,60	0,90	0,90	0,40	0,20	0,50	0,20	0,10	0,90	0,00	0,40	0,00	0,00	0,60	0,00	0,20	0,10	0,20	0,20	0,10	0,10	0,70	0,10	0,60	0,30	0,10	0,20	0,00	0,10	0,70	0,34
	P15	0,40	0,80	0,73	0,47	0,27	0,53	0,13	0,27	0,73	0,00	0,27	0,00	0,00	0,40	0,00	0,20	0,20	0,13	0,13	0,07	0,13	0,67	0,20	0,47	0,33	0,07	0,13	0,00	0,07	0,47	0,30

Figure 4.1 Comparison of various metrics for Medline Abstracts task

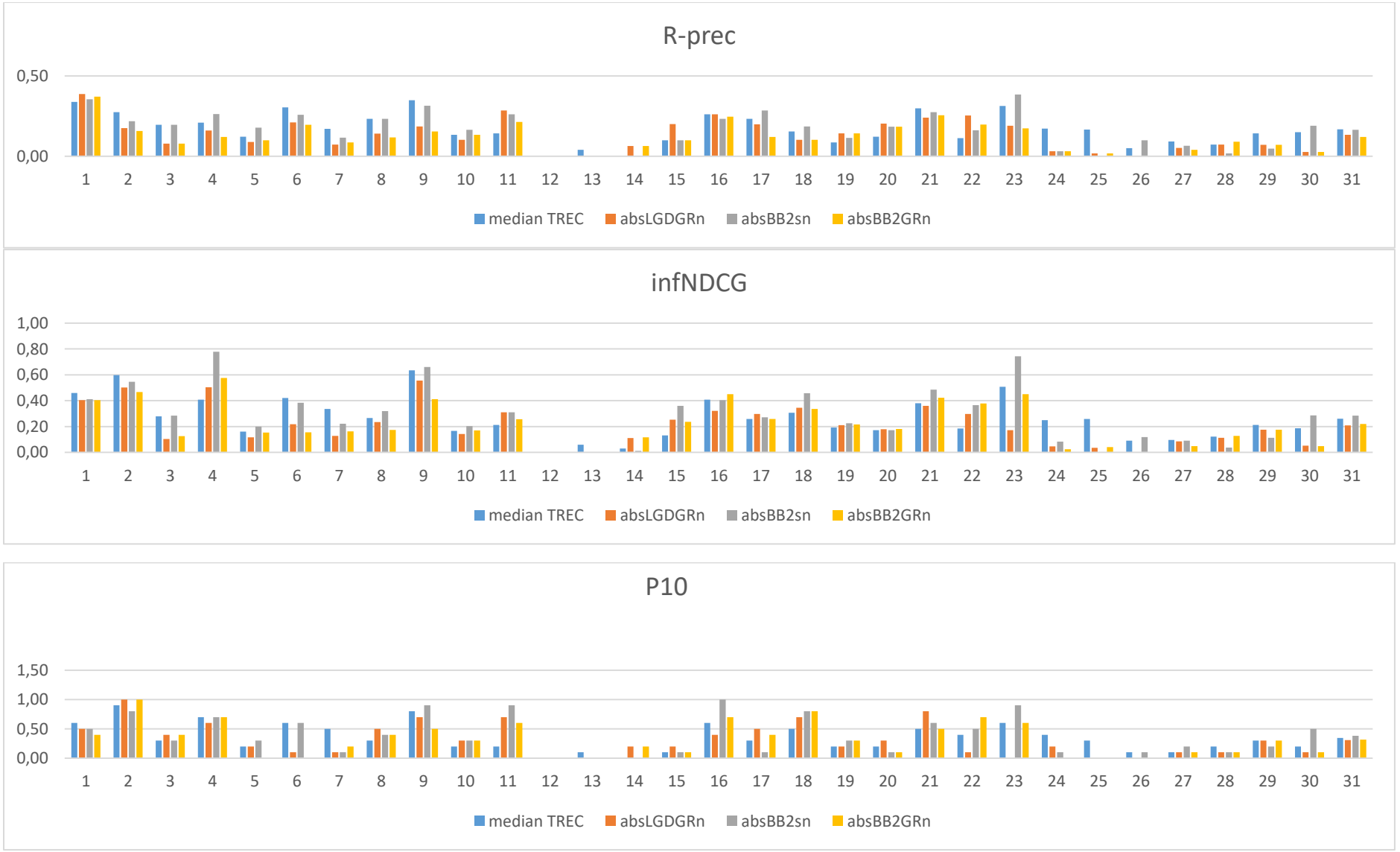
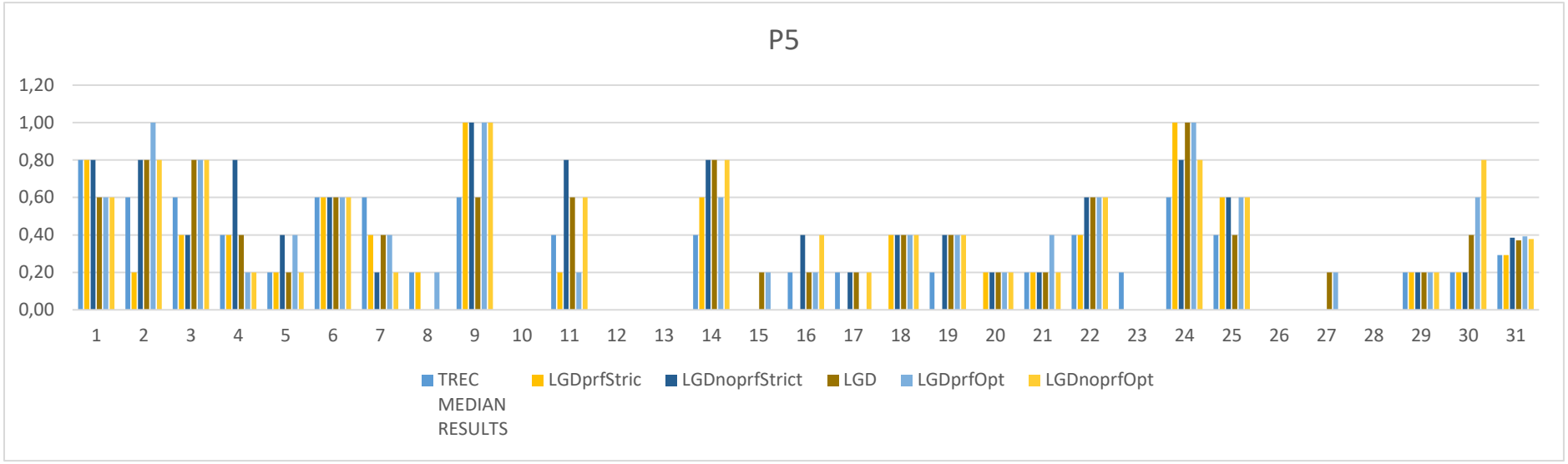
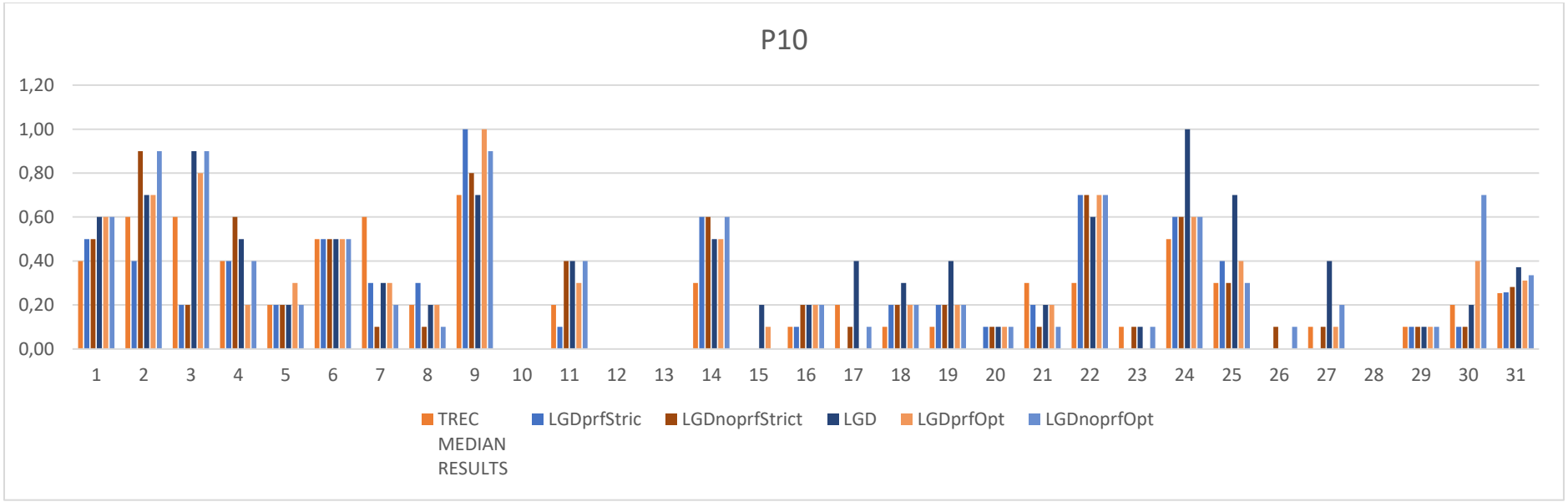


Figure 4.2 Comparison of P5 and P10 for Clinical Trials task





5. Evaluation

In the evaluation, we focus mainly on the results achieved for the Clinical Trials test set, as it was a main focus of our contribution. We attach the evaluation of results generated with the Medical Abstracts test set to the figures. The main purpose of the evaluation is the research on the poor performance of strict queries (used in LGDStrict and LGDprfStrict runs).

5.1 Evaluation framework

We use a precision and recall framework to evaluate the method. As we are operating on a partially annotated corpus, we provide dedicated definitions of recall and precision. We believe that our definitions follow the spirit of an original, commonly used precision and recall framework.

A traditional TREC evaluation methodology suffers from serious distortion of results when a number of relevant documents is smaller than @k. Beyond the number of relevant documents the results are padded with zeroes. This does not change a ranking, but caused the loss of normalization between questions, and in our opinion distorts averages. We propose a method that rectifies this shortcoming.

The annotated corpus provides a set of relevant and irrelevant documents with the annotation of relevancy for each document with regard to a given query. The result sets consist of retrieved documents with a score related to each document and a specific query. Let us define ' a number of true positives at *k* retrieved documents' (*tp@k*) as a number of retrieved documents, which are annotated as relevant. Analogically, 'a number of false positives at *k* retrieved documents' (*fp@k*) is defined as a number of retrieved documents which are annotated as irrelevant. 'A number of true negatives at *k* retrieved documents' (*tn*@k) is a number of documents, which are annotated as relevant and are not retrieved. We define precision(*prec*) as a number of relevant and retrieved documents divided by a number of retrieved and annotated (either as relevant or irrelevant) documents.

$$prec@k = \frac{tp@k}{tp@k + fp@k} \tag{1}$$

Similarly to the precision measure, we define recall as a number of retrieved and relevant documents divided by a number of all relevant documents.

$$recall@k = \frac{tp@k}{tp@k + tn@k} \tag{2}$$

It should be noted that, with this definition of recall, the ideal score might not be equal to one, as the number of relevant documents might be higher than *k*. In that case the ideal number of retrieved and relevant documents should be equal to the number of retrieved documents. We can apply this amendment to the formula by applying a rectifier to the number of retrieved and annotated documents.

$$recall@k = \frac{tp@k}{\min(tp@k + tn@k, k)} \tag{3}$$

While absolute values produced by (2) and (3) are different, rankings of methods obtained with both formulas are equal. As we believe there is little to no difference between both methods of evaluation, we are using

equation (2). The same thinking can be applied to the precision formula. Given the definitions of precision and recall we define the evaluation metric F1. We use a classical definition given by (4).

$$F1@k = 2\frac{prec@k \cdot recall@k}{prec@k + recall@k}, \tag{4}$$

although strictly speaking we should use modified precision and recall in Eq. (4). We do not do this in this work since F1 is not a primary measure of evaluation in TREC.

We expect the F1 measure within a wide range of a parameter *k* to highlight the details of invalid behavior of restricting queries.

5.2 Evaluation Setup

We export annotated sets of documents and queries, as well as the result sets to an SQL database. Architecture of the database is illustrated in the Figure 5.1. Within the database, we join the Annotations and Results tables. Based on the joined table, we construct a set of SQL queries, which allow us to retrieve *tp@k*, *tn@k* and *fp@k* measures. Finally, we define a procedure, which generates a view of retrieved values for *k* of 10, 100 and 1000 for each query.

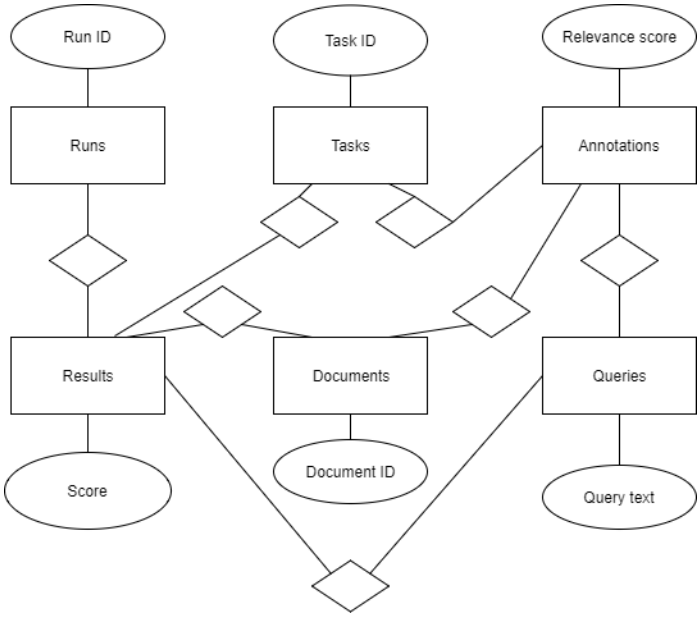


Figure 5.1 Evaluation database architecture

5.3 Evaluation Results

We present evaluation results for all three of the evaluation measures – recall, precision and F1 measure at 10, 100 and 1000 retrieved documents. Figure 5.2 shows, that our extension performs well within a large number of retrieved documents due to the relatively large increase in precision. For 1000 retrieved documents, we observe a drop of recall. That behavior is intuitive, with strict queries, we expect to retrieve a smaller number of relevant documents, at the same time, the results to be precise. The semi-strict queries (with the optional appearance of the gene name within the document) are between the baseline queries and the strict queries. The performance worsens as the number of retrieved documents gets smaller. We hoped to achieve a similar result regardless of a

parameter *k*. Figure 5.3 illustrates the behavior of restricting queries at 100 retrieved documents. Strict queries are still applicable at that point, however we already observe a drop in the F1 measure. With a very small number of retrieved documents, strict queries perform worse than simple queries, that is presented in Figure 5.4. Strict queries improve the quality of documents within the tail of the distribution. Simple, non-restricted queries retrieve equally precise documents in the head of the distribution, while preserving the higher value of recall measure.

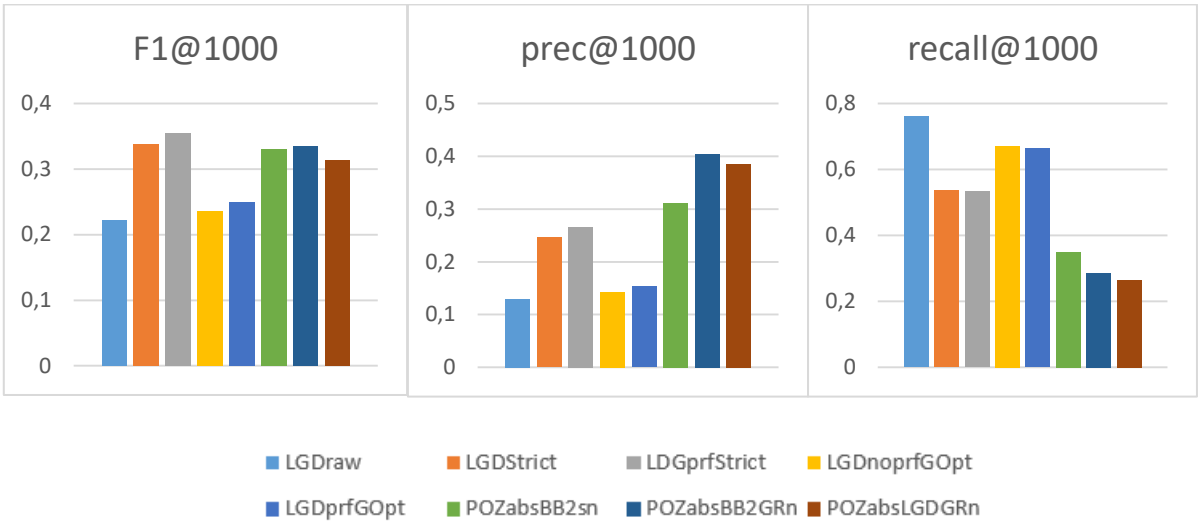


Figure 5.2 F1, precision and recall measures for 1000 retrieved documents

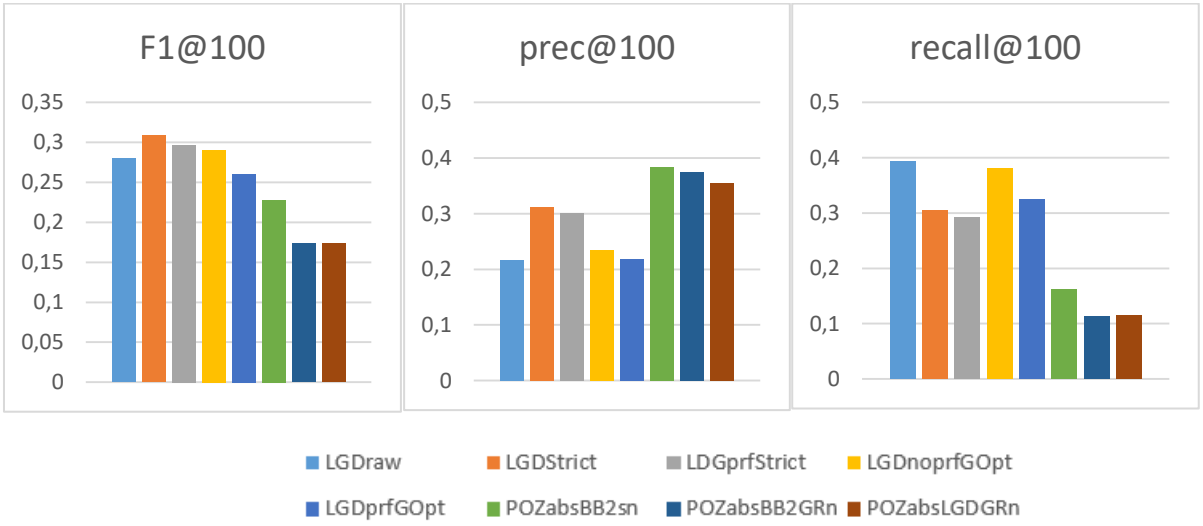


Figure 5.3 F1, precision and recall measures for 100 retrieved documents

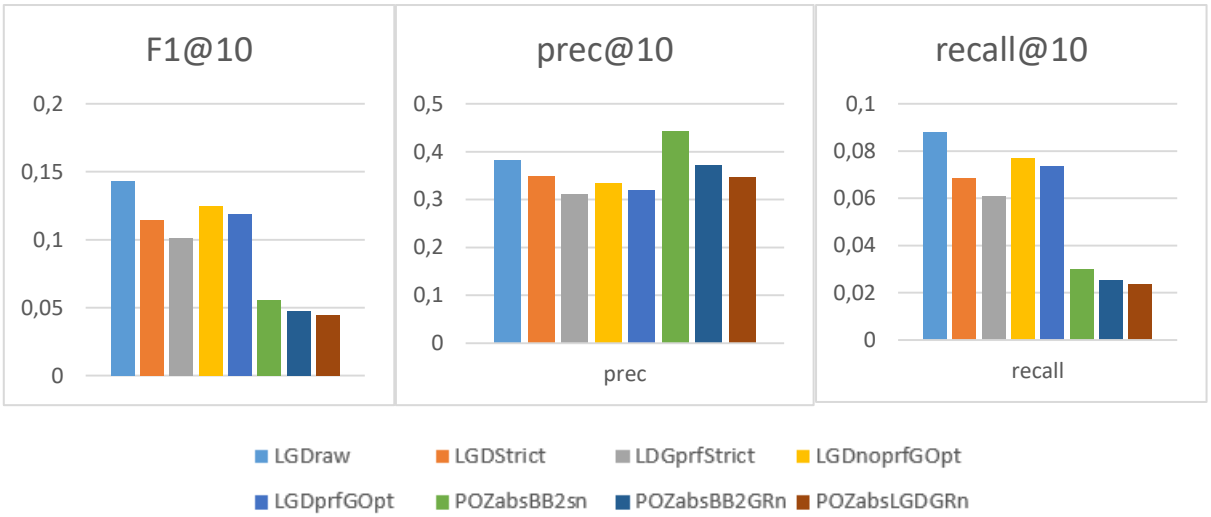


Figure 5.4 F1, precision and recall measures for 10 retrieved documents

6 Conclusions and future work

Better results were generated for Clinical Trials task (P5, P10 and P15 better than median by averagely 0.07, 0.06 and 0.05) than for Scientific Abstracts (infNDCG and P10 were better than median by 0.03 only in one BB2 run). The difference may be due to the fact, that in the Scientific Abstracts task we did not check if the documents were describing patients with correct age and gender, as this data was hard to be extracted from unstructured text. Surprisingly, more strict queries (requiring the presence of all terms in relevant document or all disease terms and any of gene terms) provided worse results than a baseline query (all query terms were treated as optional). We provide an extended evaluation of that problem. It turns out, that restricted queries enhance the tail, while worsening the head of retrieval distribution. With small numbers of 5, 10 and 15 retrieved documents strict queries perform worse.

A traditional TREC evaluation methodology suffers from serious distortion of results when a number of relevant documents is smaller than @k. Beyond the number of relevant documents the results are padded with zeroes. This does not change a ranking, but caused the loss of normalization between questions., and in our opinion distorts averages. We propose a method that rectifies this shortcoming, and facilitates comparison of methods when analyzing averages.

References

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 bioCADDIE: BIOCADDIE 2016 DATASET RETRIEVAL CHALLENGE https://biocaddie.org/biocaddie2016-dataset-retrieval-challenge, access: 21 march 2017

Cohen,T., Roberts, K., Gururaj, A., Chen, X., Pournejati, S., Hersh, W.R., Demner-Fushman, D., OhnoMachado, L., Xu,H. (2017) A Publicly Available Benchmark for Biomedical Dataset Retrieval: The Reference Standard for the 2016 bioCADDIE Dataset Retrieval Challenge. Database (Oxford).

Cieślewicz, A., Dutkiewicz, J., Jędrzejek,C., (2017) Baseline and extensions approach to information retrieval of complex medical data: Poznan Contribution to bioCADDIE 2016, Database (Oxford)., in print Clinchant,S., Gaussier,E.,

(2010) Information-based models for ad hoc IR. In SIGIR'10, 234-241. Dutkiewicz, J., Jędrzejek,C., Frąckowiak,M. and Werda,P. (2016) PUT Contribution to TREC CDS 2016, The Twenty-Fifth Text REtrieval Conference (TREC 2016) Proceedings, http://trec.nist.gov/pubs/trec25/papers/IAII_PUT-CL.pdf

Dutkiewicz, J., Jędrzejek,C., (2018) Modeling Similarity Measure for to Question Answering with Vector Space Models, submitted to IJCAI 2017. Faruqui M., Dodge J., Jauhar S. K., Dyer C., Hovy E. H., Smith N. A., (2015) Retrofitting Word Vectors to Semantic Lexicons. HLT-NAACL 1606-1615 MeSH database: https://www.nlm.nih.gov/mesh/download_mesh.html, Access: 21 March 2017

Mikolov, T., Sutskever, I., Chen, K., Corrado, Corrado, G.S., and Dean, (2013), J. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 Lu Z.., Won K. W., and Wilbur W. J., (2009) Evaluation of Query Expansion Using MeSH in PubMed, Inf Retr Boston. 12(1): 69–80. Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees E., and Hersh W. (2016), State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. Inf. Retr. 19, 1-2 (April 2016), pp. 19: 113-148.

Roberts, K., Gururaj, A., Chen, X., Pournejati, S., Hersh, W.R., Demner-Fushman, D., Ohno-Machado, L., Cohen, T., Xu, H. (2017) Information Retrieval for Biomedical Datasets: The 2016 bioCADDIE Dataset Retrieval Challenge. Database (Oxford).

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. Terrier IR Platform www.terrier.org 26 Oct 2016 TREC-CDS, http://www.trec-cds.org TREC Precision Medicine / Clinical Decision Support Track (2017), http://trec-cds.appspot.com/2017.html