Empirical research on medical information retrieval

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

[1]Instytut Automatyki, Robotyki i Inżynierii Informatycznej, WE PP
[2]Department of Clinical Pharmacology, Poznan University of Medical Sciences, Poznan, Poland

December 11, 2018

Goals and Experience in general

- We are interested in extraction of meaning for biomedical, and legal texts, and for detection of malicious code from asm and binary code in academic and industrial settings.
- We participated in TREC CDS 2016[6], TREC PM 2017 track[2] and in bioCADDIE 2016[3].
- At the beginning zero competence in precision medicine
- In 2017 A. Cieslewicz (Ph.D. in molecular biology and IT engineer) joined the team

Thesis, goal and scope of work

- Scope of work : information retrieval for biomedical documents
- **Information retrieval (IR)** is the activity of obtaining information system resources relevant to an information need from a collection of information resources.
- Thesis 1: The way current challenges are organized and evaluated are not reliable, in a sense they distort results (and prevent full evaluation of methods).
- Thesis 2: Many corrections are needed to improve baseline models, some without full explanations.
- Goal: to design and validate an IR system fully internationally competitive for biomedical documents.
- Work is based on results of contemporary challenges TREC PM 2018, bioCADDIE 2016

Subject of research

- Biomedical information retrieval
  - Retrieval from articles and abstracts TREC PM
  - Retrieval from snippets extracted from specialized databases bioCADDIE 2016
- Query expansion using word embedding
- Issue of evaluation from incomplete data (in progress)
- Novel method of word embedding improvement and ranking
- Autoencoder method of query expansion (in progress)

List of publications

1. Jakub Dutkiewicz and Czesław Jędrzejek Comparison of Paragram and Glove Results for Similarity Benchmarks, in 11-th edition of International Conference on Multimedia & Network Information Systems (MISSI 2018), indexed in Web of Science pp. 236-248 3 following papers in the same series
2. Jakub Dutkiewicz and Czesław Jędrzejek Calculating Optimal Queries from the Query Relevance File, MISSI 2018, pp 249-259 3.
3. Anna Zdrojewska, Jakub Dutkiewicz and Czesław Jędrzejek Comparison of the Novel Classification Methods on the Reuters-21578 Corpus, MISSI 2018, p. 290-302 4.
4. Artur Cieslewicz, Jakub Dutkiewicz, Czeslaw Jedrzejek: Baseline and extensions approach to information retrieval of complex medical data: Poznan approach to the bioCADDIE 2016. Database 2018: bax103 (2018). **40 points**
5. Artur Cieslewicz, Jakub Dutkiewicz, Czeslaw Jedrzejek: POZNAN Contribution to TREC PM 2017. TREC 2017
6. Jakub Dutkiewicz, Czeslaw Jedrzejek, Michal Frackowiak, Pawel Werda : PUT Contribution to TREC CDS 2016. TREC 2016
7. Michal Frackowiak, Jakub Dutkiewicz, Czeslaw Jedrzejek, Marek Retinger, Pawel Werda : Query Answering to IQ Test Questions Using Word Embedding. MISSI 2016: 283-294 6,
8. Jakub Dutkiewicz, Maciej Falkowski , Maciej Nowak , Czeslaw Jedrzejek: Semantic Extraction with Use of Frames. PolTAL 2014 : 208-215

Text REtrieval Conference (TREC) https://trec.nist.gov/

The Text REtrieval Conference is an ongoing series of workshops focusing on a list of different information retrieval research areas, or tracks. It is co-sponsored by the National Institute of Standards and Technology and began in 1992. Its purpose is to support and encourage research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies and to increase the speed of lab-to-product transfer of technology.
Each track has a challenge wherein NIST provides participating groups with data sets and test problems. Depending on track, **test problems might be questions, topics, or target extractable features**. Uniform scoring is performed so the systems can be fairly evaluated.

Topics constituting queries TREC PM 2018 close to TREC PM 2017

Disease: melanoma
Variant: BRAF (V600E)
Demographic: 64-year-old male

Disease: melanoma
Variant: no tumor infiltrating lymphocytes
Demographic: 74-year-old male

Disease: gastric cancer
Variant: EGFR
Demographic: 60-year-old female

Disease: papillary thyroid carcinoma
Variant: NTRK1
Demographic: 46-year-old male

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

bioCaddie and TREC text corpora

- **bioCaddie**:
  - 788 992 datasets in 20 repositories,
  - heterogeneous data,
  - XML and JSON formats
  - combination of structured and unstructured data
- **TREC Abstracts**
  - Snapshot of PubMed (or PMC aka Pub Med Central) database
  - Abstracts of scientific publications
  - ~1 000 000 documents (varies yearly)
- **TREC Clinical Trials**
  - Snapshot of Clinical Trials database
  - Combination of structured and unstructured data
  - Semistructured description of conducted clinical trial
  - ~100 000 documents

The bioCaddie document

```
<DOC>
<DOCNO>500000</DOCNO>
<TITLE>A375R_RPL10a_vivo__Ronly_vem10d_rep2</TITLE>
<REPOSITORY>geo_022216</REPOSITORY>
<METADATA>
{"dataItem":
        { "Type":"NA",
          "source_name": "melanoma",
          "description": "NA",
          "title": "A375R_RPL10a_vivo__Ronly_vem10d_rep2",
          "series": "GSE64741",
          "geo_accession": "GSM1579183",
          "platform": ["GPL11154"],
          "citations": 0,
          "link": "SRASRX832388",
          "assays": "NA",
          "entry_type": "Sample",
          "organism": "Homo sapiens",
          "ID": "301579183"}}
</METADATA>
 </DOC>
```

The TREC Abstract document

```
<Article PubModel="Print">
        <Journal>
                <ISSN IssnType="Print">0374-5600</ISSN>
                <JournalIssue CitedMedium="Print">
                <Volume>33</Volume>
                <Issue>1</Issue>
                <PubDate>
                        <Year>1991</Year>
                        <Month>Feb</Month>
                </PubDate>
                </JournalIssue>
                <Title>Acta paediatrica Japonica : Overseas edition</Title>
                <ISOAbbreviation>Acta Paediatr Jpn</ISOAbbreviation>
        </Journal>
        <ArticleTitle>Association of neonatal thrombocytopenia and maternal anti-HLA antibodies.</ArticleTitle>
        <Pagination> <MedlinePgn>71-6</MedlinePgn> </Pagination>
        <Abstract> <AbstractText>In order to evaluate the influence of maternal anti-HLA antibody on neonatal
        thrombocytopenia, clinical features and maternal anti-HLA antibody of three groups of infants...
        </Abstract>
</Article>
```

The TREC Clinical Trials document

```
<?xml version="1.0" encoding="UTF-8"?>
<clinical_study rank="240642">
  <!-- This xml conforms to an XML Schema at:
    https://clinicaltrials.gov/ct2/html/images/info/public.xsd -->
  <required_header>
    <download_date>ClinicalTrials.gov processed this data on April 10, 2017</download_date>
    <link_text>Link to the current ClinicalTrials.gov record.</link_text>
    <url>https://clinicaltrials.gov/show/NCT00000485</url>
  </required_header>
  <id_info>
    <org_study_id>4</org_study_id>
    <nct_id>NCT00000485</nct_id>
  </id_info>
  <brief_title>Hypertension Detection and Follow-up Program (HDFP)</brief_title>
  <sponsors>
    <lead_sponsor>
      <agency>National Heart, Lung, and Blood Institute (NHLBI)</agency>
      <agency_class>NIH</agency_class>
    </lead_sponsor>
  </sponsors>
  <source>National Heart, Lung, and Blood Institute (NHLBI)</source>
  <brief_summary>
    <textblock>
      To determine the effectiveness of systematic, sustained, antihypertensive therapy in
      reducing morbidity and mortality from hypertension in a wide spectrum of persons with
      elevated blood pressure in 14 communities. During its course, the trial also obtained a
      direct measure of the prevalence, severity, and treatment status of representative white and
      black populations with high blood pressure in these 14 communities, and obtained an estimate
      of the extent of attainable reduction of complications of high blood pressure by an
      organized screening and blood pressure management program.
    </textblock>
  </brief_summary>
  <detailed_description>
    <textblock>
      BACKGROUND:

      Published data from the Veterans Administration Cooperative Study of Hypertension
      demonstrated that reduction in morbidity and mortality could be attained by treating men
```

Manual Judgements

		Total	PM	Disease	Gene	Demo.	Other
2017	Articles	22,642	9,274	5,422	2,874	8,394	9,109
	Trials	13,441	3,961	1,816	1,729	3,597	3,850
2018	Articles	22,429	9,224	7,083	4,927	8,546	8,996
	Trials	14,188	5,814	3,422	2,150	5,438	5,750

Number of PM docs significantly larger than a Gene containing docs so a keyword seach not effective

Manual Judgments

		Total	Def. Relevant	Partial. Relevant	Not Relevant
2017	Articles	22,642	2,022	1,853	18,767
	Trials	13,441	436	735	12,270
2018	Articles	22,429	3,436	2,117	16,876
	Trials	14,188	1,172	872	12,144

Number of judged documents is significantly lower than total number of documents.

Evaluation measures

- $AP = \sum_{k=1}^{n} P(k)\Delta rel(k) = \frac{\sum_{k=1}^{n} P(k)rel(k)}{\#relevant}$
- $NDCG = \frac{rel(1)+\sum_{k=2}^{n}\frac{rel(k)}{log_2(k)}}{IDCG} = \frac{rel(1)+\sum_{k=2}^{n}\frac{rel(k)}{log_2(k)}}{1+\sum_{k=2}^{n}\frac{1}{log_2(k)}}$
- infAP - inferred version of AP
- infNDCG - inferred version of NDCG

Evaluation measures[11, 12]

Inferred AP

[Yilmaz and Aslam CIKM07]

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

$$\widehat{PC}_1 = \frac{1}{1} \cdot 1 = 1$$

$$\widehat{PC}_3 = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{2} = 0.625$$

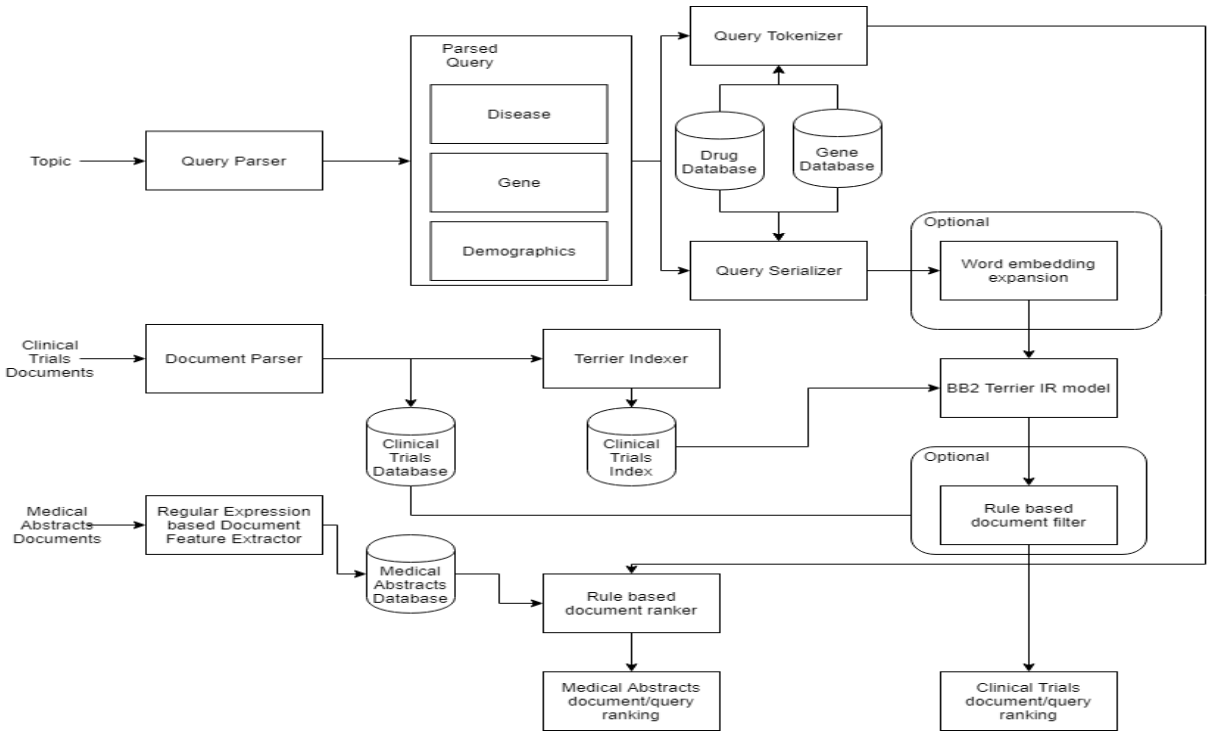
$$\widehat{PC}_9 = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \frac{2}{4} = 0.5555$$

$$\text{infAP} = \frac{1 + 0.625 + 0.5555}{3} = 0.7268$$

Evaluation measures

- $AP = \sum_{k=1}^{n} P(k)\Delta rel(k) = \frac{\sum_{k=1}^{n} P(k)rel(k)}{\#relevant}$
- $NDCG = \frac{rel(1)+\sum_{k=2}^{n} \frac{rel(k)}{log_2(k)}}{IDCG} = \frac{rel(1)+\sum_{k=2}^{n} \frac{rel(k)}{log_2(k)}}{1+\sum_{k=2}^{n} \frac{1}{log_2(k)}}$
- infAP - inferred version of AP
- infNDCG - inferred version of NDCG
- P@k - precision at k retrieved documents
- NDCG@k - normalized discounted cumulative gain at k retrieved documents

TREC Poznan Retrieval methodology setup



Assumptions of our approach

- No relations determinantion
- Clinical trials always provide evidence for treatment - wrong
- Concentrate on new „baseline" for medical tracks (parametrize the method):
  - Terrier engine DFR options BB2 or LGD
  - Word embedding [1]
  - Weights for QE terms

TREC PM 2018 Poznan runs

Terrier runs were carried out, using BB2 as ranking function:

- BB2_simple_noprf: simple query was put as input for terrier
- 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) BB2_variant_noprf: variant query was put as input for terrier
- BB2_variant_w2v_prf: variant query expanded with word2vec was put as input for terrier; additional expansion was carried out using terrier PRF

TREC PM 2018 Results: Clinical Trials

	infNDCG
hpi-dhc	0.5545
MSIIP	0.5503
ims_unipd	0.5395
UCAS	0.5347
udel_fang	0.5057
NOVASearch	0.4992
Poznan	0.4894
UTDHLTRI	0.4794
RSA_DSC	0.4743
IRIT	0.4736

	P@10
MSIIP	0.6260
ims_unipd	0.5660
Poznan	0.5580
NOVASearch	0.5520
RSA_DSC	0.5480
UCAS	0.5460
hpi-dsc	0.5400
UTDHLTRI	0.5380
udel_fang	0.5240
InfoLabPM	0.5240

	R-prec
MSIIP	0.4294
imi_unipd	0.4128
Poznan	0.4101
hpi-dhc	0.4081
UCAS	0.4005
udel_fang	0.3967
NOVASearch	0.3931
UTDHLTRI	0.3920
RSA_DSC	0.3721
IRIT	0.3658

Results for various options for the provided Clinical Trials data runs - infNDCG

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

Results for various options for the provided Clinical Trials data runs - infNDCG

topic	sq_nprf	vq_noprf	vqw2vprf	sqw2vprf	sq	trec_best	trec_median
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

Variation of measures for each bioCADDIE question: CDJ Database (2018)

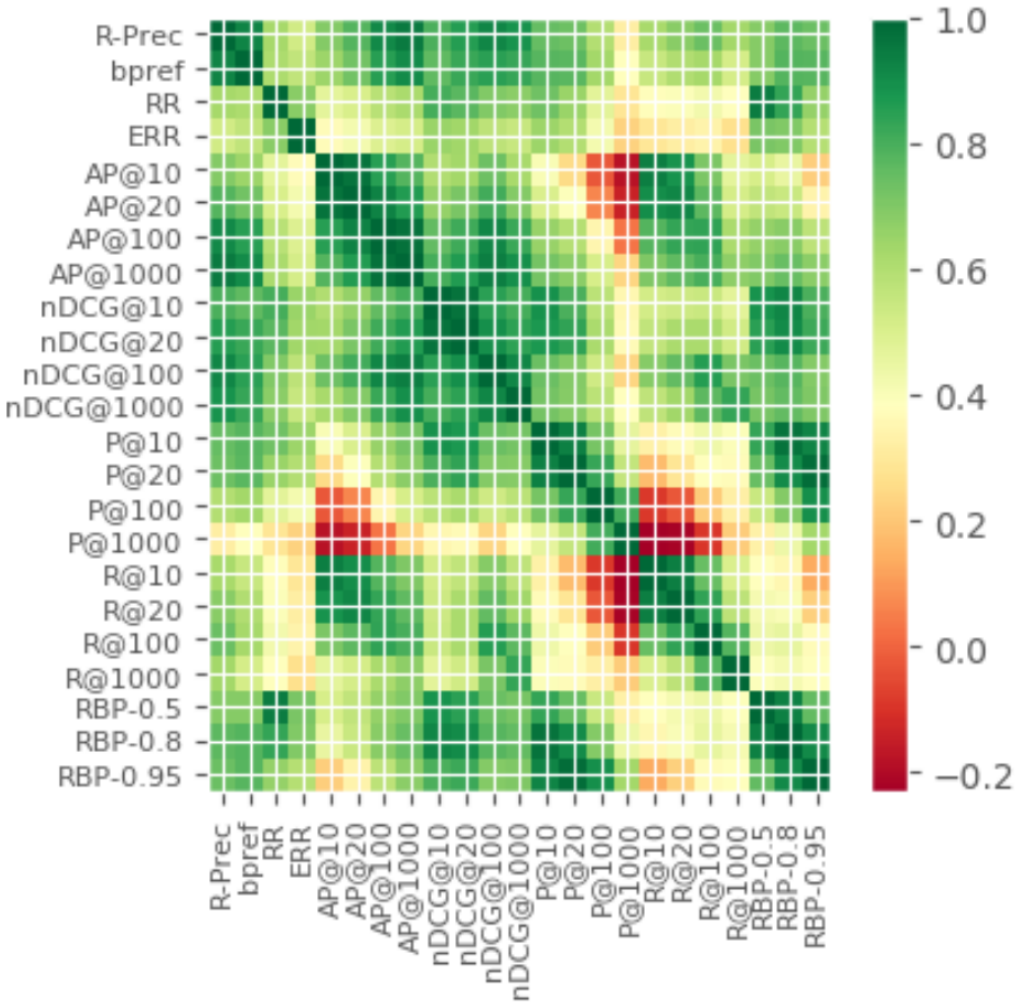
Query Number	infAP	infNDCG	P@10(partial)
1	0.4217	0.6504	0.9000
2	0.3933	0.3338	0.8000
3	0.5832	0.6898	0.9000
4	0.6999	0.5177	1.0000
5	0.1620	0.2897	0.4000
6	0.3256	0.4938	1.0000
7	0.1931	0.6197	0.2500
8	0.0856	0.4547	0.3000
9	0.2207	0.2607	0.8000
10	0.1186	0.1961	0.5000
11	0.6373	0.3402	1.0000
12	0.5860	0.4011	0.9000
13	0.3171	0.2919	0.9000
14	0.7005	0.3300	0.9000
15	0.5228	0.9384	1.0000
Average	0.3978	0.4539	0.7700

Poznan consortium results as submitted for the challenge

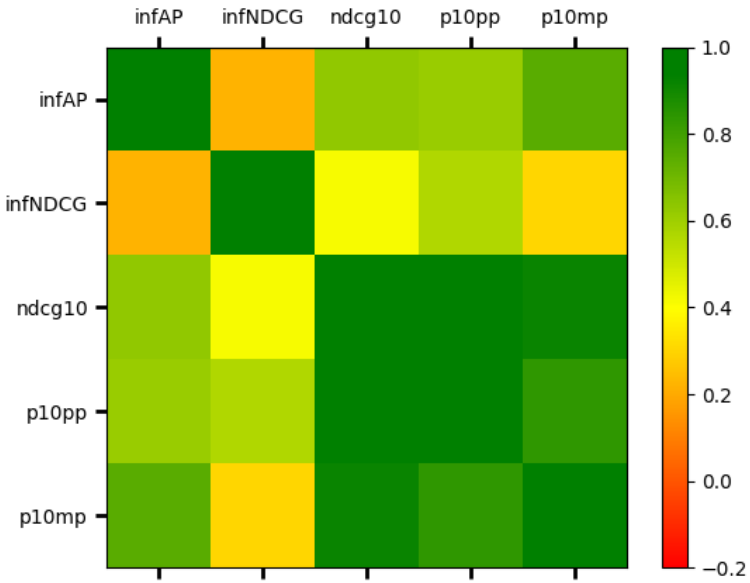
vs. the best participant results, for a given evaluation measure (in bold font). The results of this work (the Poznan consortium) are shown in italic.

Group	Submission	infAP	infNDCG	NDCG@10	P@10 (+partial)	P@10 (-partial)
IAII_PUT	biocaddie dphresults.txt	0.0876	0.3580	0.4265	0.5333	0.1600
UCSD paper	all PSD	0.2792	0.4980	0.612	0.7600	0.3267
UCSD challenge	armyofucsdgrads-3.txt	0.1468	**0.5132**	0.5303	0.7133	0.2400
SIBTex	sibtex-5_0.txt	0.3664	0.4188	0.6271	0.7533	0.3467
Elsevier	elsevier4.txt	0.3049	0.4368	**0.6861**	**0.8267**	**0.4267**
Poznan (paper)	*LGD word2vec and Terrier Rocchio*	***0.3978***	*0.4539*	*0.6375*	*0.7700*	*0.4000*

Pearson correlation between measures based on TREC historical data[7]



Pearson correlation between measures for bioCADDIE challenge (indNDCG based on 1000 documents)



Statistics of evaluated documents

Rank/Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	sum
-1	9799	9961	8597	6159	8084	8557	8145	7683	9749	7047	11343	5778	7326	7541	6852	122621
0	994	1029	657	1127	1376	872	1622	1657	1084	993	1520	732	1146	586	908	16303
1	460	33	580	256	80	361	94	72	109	150	50	100	119	121	484	3069
2	177	6	15	23	4	31	0	5	30	94	184	26	88	129	0	812
relevant	637	39	595	279	84	392	94	77	139	244	234	126	207	250	484	3881
labeled	1631	1068	1252	1406	1460	1264	1716	1734	1223	1237	1754	858	1353	836	1392	20184
relevant/labeled	0,390	0,037	0,475	0,198	0,058	0,310	0,055	0,044	0,114	0,197	0,133	0,147	0,153	0,299	0,3478	0,192
labeled/all	0.143	0.097	0.127	0.186	0.153	0.129	0.174	0.184	0.111	0.149	0.134	0.129	0.156	0.100	0.169	0.141

Number of annotated documents in the BioCaddie challenge and ratios between number of labeled, relevant and all documents.

BioCaddie queries

No.	Query
1	Find protein sequencing data related to bacterial chemotaxis across all databases
2	Search for data of all types related to MIP-2 gene related to biliary atresia across all databases
3	Search for all data types related to gene TP53INP1 in relation to p53 activation across all databases
4	Find all data types related to inflammation during oxidative stress in human hepatic cells across all databases
5	Search for gene expression and genetic deletion data that mention CD69 in memory augmentation studies across all databases
6	Search for data of all types related to the LDLR gene related to cardiovascular disease across all databases
7	Search for gene expression datasets on photo transduction and regulation of calcium in blind D melanogaster
8	Search for proteomic data related to regulation of calcium in blind D melanogaster
9	Search for data of all types related to the ob gene in obese M musculus across all databases
10	Search for data of all types related to energy metabolism in obese M musculus
11	Search for all data for the HTT gene related to Huntington's disease across all databases
12	Search for data on neural brain tissue in transgenic mice related to Huntington's disease
13	Search for all data on the SNCA gene related to Parkinson's disease across all databases
14	Search for data on nerve cells in the substantia nigra in mice across all databases
15	Find data on the NF-κB signaling pathway in MG (Myasthenia gravis) patients

Distilled queries

1. protein sequencing bacterial chemotaxis
2. MIP-2 gene biliary atresia
3. gene TP53INP1 p53 activation
4. inflammation oxidative stress human hepatic cells
5. gene expression genetic deletion CD69 memory augmentation
6. LDLR gene cardiovascular disease
7. gene expression photo transduction regulation of calcium blind D melanogaster
8. proteomic regulation of calcium blind D melanogaster
9. ob gene obese M musculus
10. energy metabolism obese M musculus
11. HTT gene Huntington disease
12. neural brain tissue transgenic mice Huntington disease
13. SNCA gene Parkinson disease
14. nerve cells substantia nigra mice
15. NF-κB signaling pathway MG Myasthenia gravis patients

Distributional hypothesis

- "a word is characterized by the company it keeps",
- Formulated by Harris in 1954,
- Basis of Statistical and Distributional Semantics,
- If two different words often appear with similar contexts they are replaceable by each other, either in document or in a query,
- Word2Vec is one of the implementations of the hypothesis.

Terminology encoder

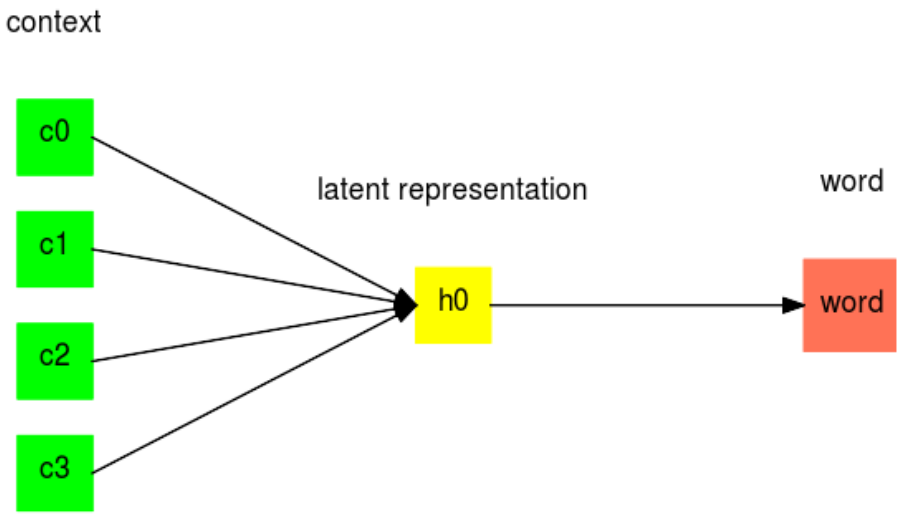


Figure: Language model architecture - CBOW encoder

Similarity between words calculated upon their latent representations

Given latent representations v1 and v2 for words w1 and w2, similarity between those word is calculated as similarity between the latent word representations[5].

$$sim(w1, w2) = sim(v1, v2)$$

Using the angular definition of distance

$$sim(w1, w2) = cos(v1, v2)$$

Using the euclidean definition of distance:

$$sim(w1, w2) = 1 - d(v1, v2)$$

If similarity between query word and a candidate word is above given threshold, the query is expanded with the candidate word. We strictly use the cosine definition of distance and a threshold of 0,8. We use two separate language models, a classic word2vec model calculated on the biocaddie text corpus and a version implemented in [1].

Weights for QE terms - unexplained

Topic terms weight =160
Expanded terms with Korhonen language model weight = 5
Expanded terms language with model from bioCADDIE weight = 1
...

- 9 ob^160 gene^160 obese^160 m^160 musculus^160 **genes^5 normal-weight^5 non-obese^5 overweight^5 obesity^5 overweight/obese^5 Mus^5** lean leptine Lep ghrelin satiety
- 10 energy^160 metabolism^160 obese^160 m^160 musculus^160 **Energy^5 metabolisms^5 normal-weight^5 non-obese^5 overweight^5 obesity^5 overweight/obese^5 Mus^5** lean
- 11 htt^160 gene^160 huntington^160 disease^160 **huntingtin^5 Htt^5 mHtt^5 polyQ^5 ataxin-3^5 genes^5**

Retrieval model

- We use the Terrier Information Retrieval implementation[10]
- The implementation follows the basic Bayesian model for retrieval

$$P(Q|D) = \prod_{q \in Q} \frac{P(D|q)P(Q)}{P(D)} \propto logP(Q) + log \sum_{q \in Q} P(D|q)$$

There are several implemented probability models for the conditionals and priors estimation.

Retrieval model

- We use various Divergence From Randomness models for the Information Retrieval task
- Divergence from randomness model examines the divergence between distribution of a word in a given document and distribution of the same word within a set of documents.
- An example DFR formula is given by a product of two divergence functions[9]

$$\sum_i I_1(\hat{p}_i^+ || \hat{p}_i) \cdot I_2(\hat{p}_i || p_i)$$

- where $\hat{p}$ is the frequency of terms in a document, $\hat{p}^+$is a frequency of the neighbouring terms in a document and p is a prior probability density function the terms in the entire set.
- If $p_i$is similar to $\hat{p}_i$ then the term occured "randomly", term is more informative, if $\hat{p}_i \gg p_i$.
- We use several DFR models, such as BB2, LGD or DPH.[4]

Calculating the Optimal Queries for a given Query Relevance file

- Documents annotated as non-relevant D^{POS} and relevant D^{NEG}
- Term usability

$$ir(t_i)_{D^{POS},D^{NEG}} = \frac{idf_{D^{POS}}(t_i)}{idf_{D^{NEG}}(t_i)} \cdot \frac{1}{|idf_{D^{NEG}} - idf_{D^{POS}}|}$$

- Term representativeness

$$s(t_i)_{D^{POS},D^{NEG}} = \frac{tf_{D^{POS}}(t_i)}{tf_{D^{NEG}}(t_i)} \cdot |tf_{D^{NEG}} - tf_{D^{POS}}|$$

- Term evaluation scores

$$score_1(t_i, Q_j) = \frac{s_{Q_j}(t_i)}{ir_{Q_j}(t_i)}$$

$$score_2(t_i, Q_j) = \frac{k_1}{ir_{Q_j}(t_i)} + k_2 \cdot s_{Q_j}(t_i)$$

The Optimal Queries - results

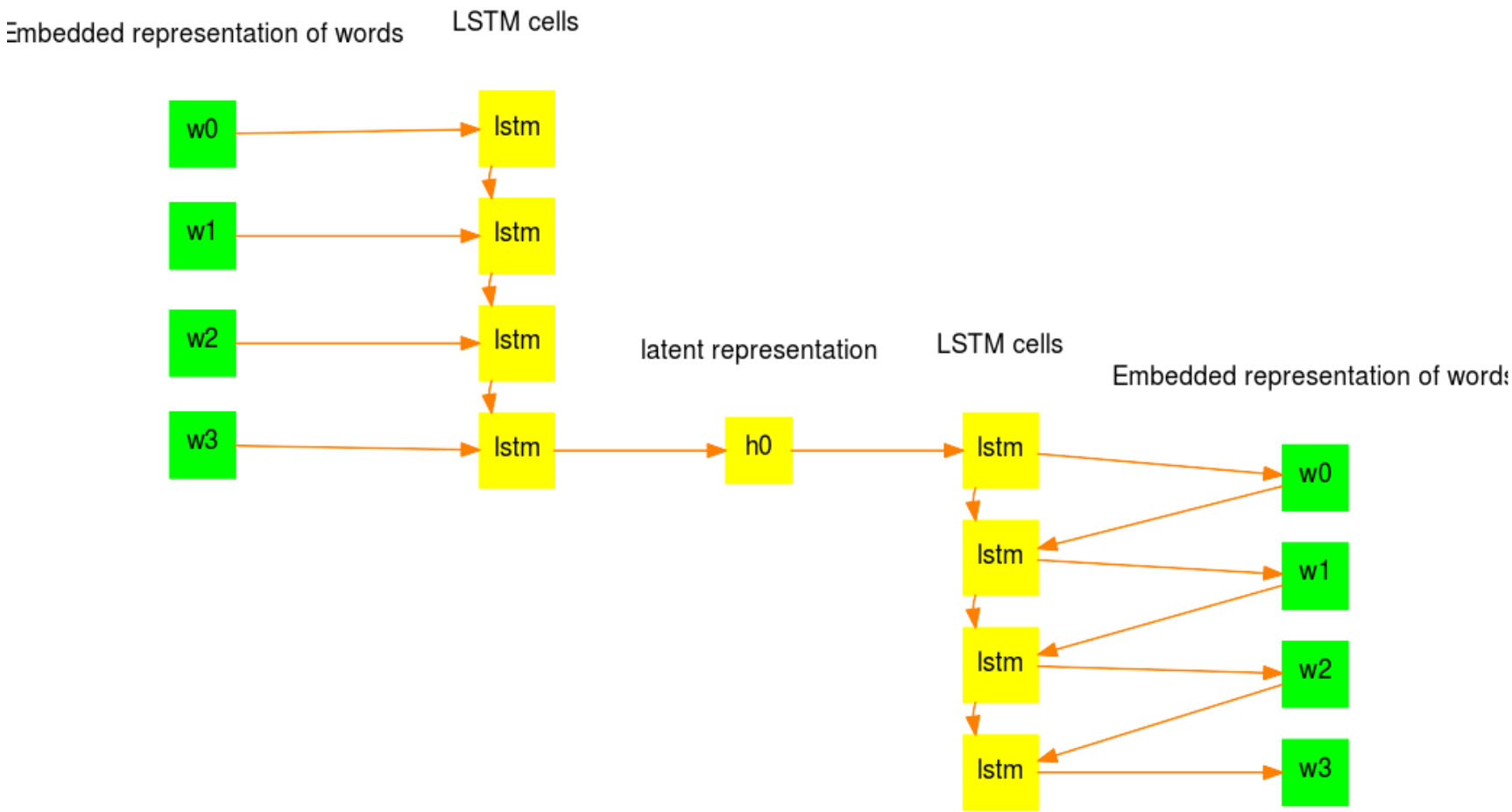
Table 7. Evaluation scores for the Optimal Queries compared to TREC results

Query Creation model	N	P@10	R-prec	infAP	infNDCG
TREC best results	-	0.4033	0.1744	0.0454	0.2815
$Score_1$	25	0.4167	0.2016	0.0829	0.3769
$Score_1$	30	0.4400	0.2024	0.0806	0.3853
$Score_1$	35	0.4567	0.2006	0.0867	0.4107
$Score_1$	40	0.5000	0.2161	0.0945	0.4060
$Score_1$	45	0.4767	0.2187	0.0973	0.4127
$Score_1$	50	0.5067	0.2296	0.0997	0.4089
$Score_1$	55	0.4867	0.2257	0.0999	0.4007
$Score_1$	60	0.5133	0.2287	0.1002	0.4057
$Score_2$ ($k_1 = 1$; $k_2 = 0$)	30	0.4500	0.2011	0.0910	0.3728
$Score_2$ ($k_1 = 10$; $k_2 = 1$)	30	0.4500	0.2006	0.0907	0.3718
$Score_2$ ($k_1 = 1$; $k_2 = 1$)	30	0.4500	0.2007	0.0904	0.3697
$Score_2$ ($k_1 = 1$; $k_2 = 10$)	30	0.4667	0.2013	0.0932	0.3809
$Score_2$ ($k_1 = 1$; $k_2 = 0$)	30	0.4667	0.2013	0.0895	0.3783
$Score_2$ ($k_1 = 1$; $k_2 = 1$)	40	0.4600	0.2038	0.0886	0.3684
$Score_2$ ($k_1 = 1$; $k_2 = 1$)	50	0.4800	0.2060	0.0950	0.3942
$Score_2$ ($k_1 = 1$; $k_2 = 1$)	60	0.4700	0.2063	0.0953	0.3887

Autoencoder for document compression

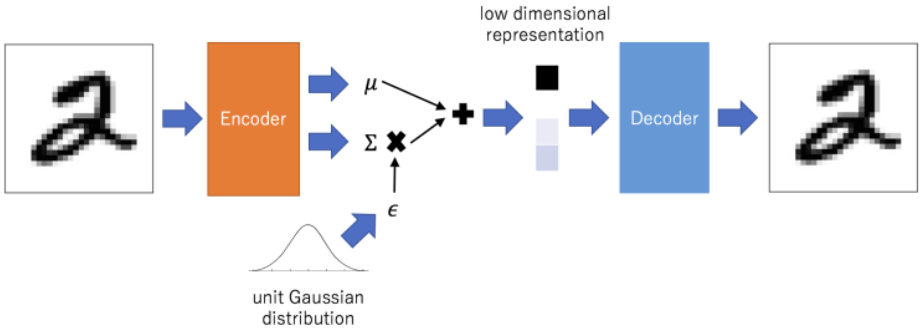
- We plan to use the Autoencoder neural network to generate latent representation of both queries and documents,
- We want to use the idea similar to the one presented in [8],
- We train the encoder on both documents and queries,
- We compare the latent representations of document to latent representations of queries, based on that comparison we pick documents which are similar to queries,
- We plan to use several types of word embeddings - specifically, a classical one-hot embedding and word embedding created with an imlpementation of Distributional Hypothesis (i.e. Paragram, Glove, Word2Vec),
- Recent works report increase of evaluation measures we are specifically interested in - NDCG and MAP[10].

Autoencoder for document compression (latent representation generation)



Variational Autoencoder like approach

- We want to optimize term usability and term representativeness for a given query
- Similarily to optimizing the mean and standard deviation in variational autoencoder



[1]

[1] http://mlexplained.com/2017/12/28/an-intuitive-explanation-of-variational-autoencoders-vaes-part-1/

Conclusions

- Despite very simplified assumption (no evidence for treatment terms and relations) our new baseline is strong
- Pseudo Relevance Feedback (PRF) makes the results worse – we do not match top documents very well
- the best result is vq_noprf option which is significantly better – (approximately 0.06-0.08 above median for evaluated measures: infNDCG, R_prec P@10)
- With a suitable word2vec method the results are better compared to query extension using Mesh and disease taxonomies
- We need to prepare data for queries from various competitions to create a reliable machine learning based implementation for information retrieval

[1] Billy Chiu, Gamal K. O. Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical NLP. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun'ichi Tsujii, editors, *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2016, Berlin, Germany, August 12, 2016*, pages 166–174. Association for Computational Linguistics, 2016.

[2] Artur Cieslewicz, Jakub Dutkiewicz, and Czeslaw Jedrzejek. POZNAN contribution to TREC PM 2017. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume Special Publication 500-324. National Institute of Standards and Technology (NIST), 2017.

[3] Artur Cieslewicz, Jakub Dutkiewicz, and Czeslaw Jedrzejek. Baseline and extensions approach to information retrieval of complex medical data: Poznan's approach to the biocaddie 2016. *Database*, 2018:bax103, 2018.

[4] Stéphane Clinchant and Eric Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for ir.

In Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz, editors, *Advances in Information Retrieval Theory*, pages 54–65, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[5] Jakub Dutkiewicz and Czeslaw Jedrzejek. Comparison of paragram and glove results for similarity benchmarks. In Kazimierz Choros, Marek Kopel, Elzbieta Kukla, and Andrzej Sieminski, editors, *Multimedia and Network Information Systems - Proceedings of the 11th International Conference MISSI 2018, Wrocław, Poland, 12-14 September 2018*, volume 833 of *Advances in Intelligent Systems and Computing*, pages 236–248. Springer, 2018.

[6] Jakub Dutkiewicz, Czeslaw Jedrzejek, Michal Frackowiak, and Pawel Werda. PUT contribution to TREC CDS 2016. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, volume Special Publication 500-321. National Institute of Standards and Technology (NIST), 2016.

[7] Mucahid Kutlu, Vivek Khetan, and Matthew Lease. Correlation and

prediction of evaluation metrics in information retrieval. *CoRR*, abs/1802.00323, 2018.

[8] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1106–1115. The Association for Computer Linguistics, 2015.

[9] D. V. Lindley. Information theory and statistics. solomon kullback. new york: John wiley and sons, inc.; london: Chapman and hall, ltd.; 1959. pp. xvii, 395. $12.50. *Journal of the American Statistical Association*, 1959.

[10] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, pages 517–519, Berlin, Heidelberg, 2005. Springer-Verlag.

[11] Emine Yilmaz, Evangelos Kanoulas, and Javed Aslam. A simple and efficient sampling method for estimating ap and ndcg. pages 603–610, 07 2008.

[12] Emine Yilmaz, Evangelos Kanoulas, and Javed Aslam. A simple and efficient sampling method for estimating ap and ndcg: Slides. 07 2008.