PUT Contribution to TREC CDS 2016

Jakub Dutkiewicz[1], Czesław Jędrzejek[1], Michał Frąckowiak[1] , Paweł Werda[1]

[1] IAiII, Poznan University of Technology

Maria Sklodowska-Curie Sq. 5, 60-965 Poznan, Poland

{jakub.dutkiewicz, czeslaw.jedrzejek}@put.poznan.pl

Abstract. This paper describes the medical information retrieval systems designed by the Poznan University of Technology of for clinical decision support CDS) which were submitted to the TREC 2016. 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 CDS 2016 is following two previous Clinical Decision Support Track contest. Its aim was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records.

1.1 Source and target documents and rules

Unlike previous years, actual electronic health record (EHR) patient records taken from admission notes in MIMIC-III were used instead of synthetic cases. The target document collection for the track was the Open Access Subset of PubMed Central (PMC). PMC is an online digital database of freely available full-text biomedical literature. Because documents are constantly being added to PMC, to ensure the consistency of the collection, for the 2016 task we obtained a snapshot of the open access subset on March 28, 2016, which contained a 1.25 million articles. The full text of each article in the open access subset is represented as an NXML file (XML encoded using the NLM Journal Archiving and Interchange Tag Library).

Detailed rules for the contest were in guidelines send to participants and in [CDS topics, 2016]. One such a rule is the following:
"System data structures (such as dictionaries, indices, thesauri, etc. whether constructed by hand or automatically) can be built using existing documents, topics, and relevance judgments, but these structures may not be modified in response to the new test topics. For example, one can't add topic words that are not in one's dictionary, nor may the system data structures be based in any way on the results of retrieving documents for the test topics and having a human look at the retrieved documents".

1.2 Experience

For the four member team at Information Systems and Technologies Division, IAiII, Poznan University of Technology this was the first participation at any TREC track. In our research we use the word embedding [Mikolov et al., 2013a] with semantic (relation) knowledge [Faruqui et.al., 2015], [Jauhar et.al., 2015], [Dutkiewicz, Jedrzejek, 2016]. We are collaborating with Poznan Medical University on mining data related stomach and melanoma cancers. We completed the work on the answer to multiple choice questions that beats the currently best literature result on ones set of test questions by 8% [Frąckowiak et al., 2016]. For another two benchmark test questions we are very close to leading literature results [ACLWEB TOEFL, 2016)], [ACLWEB ESL, 2016)]. We hoped that this experience would be transferable for Clinical Decision Support track tasks. For the lack of time and resources we were restricted to limited number of options. We did not preprocess the target Open Access Subset of PubMed Central (PMC).

2 Related work

2.1 Baseline system

Usually, in TREC research available indexing packages are used, the most popular being Terrier or Lucene systems. In the recent work [Lin, 2016] several leading systems were evaluated within the Open-Source IR Reproducibility Challenge for the Gov2 test collection. Among the options was Terrier 4,0 with DPH ranking function, which is a hypergeometric parameter-free model from the Divergence from Randomness family of functions. The query expansion version - the "DPH + Bo1 QE" uses the pseudo-relevance feedback (PRF), which is known to find potentially relevant terms by first querying the index and looking for new terms in high-ranking documents. Specifically, from which 10 terms are added from 3 pseudo-relevance feedback (PRF) documents.

[Lin et al., 2016] found that the "DPH + Bo1 QE" run of Terrier 4.0 was statistically significantly better than all other runs including Terrier's BM25 run, with all other differences were not significant.

2.2 Query expansion

Expanding queries by adding potentially relevant terms is a common practice in improving relevance in IR systems. The idea is to add synonyms, and other similar terms to query terms. 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, SNOMED-CT, ICD-10, WordNet, and Wikipedia are used. Generally, the result of lexicon type of expansion is positive. The second category is word embedding. [Goodwin and Harabagiu, 2014] used the Skip-gram word2vec method for query expansion with negative effect compared to baseline.

[Wang and Fang, 2014] reported their best result on a method that utilizes both PRF and UMLS- based expansion, and it appears that UMLS contributed more to the performance.

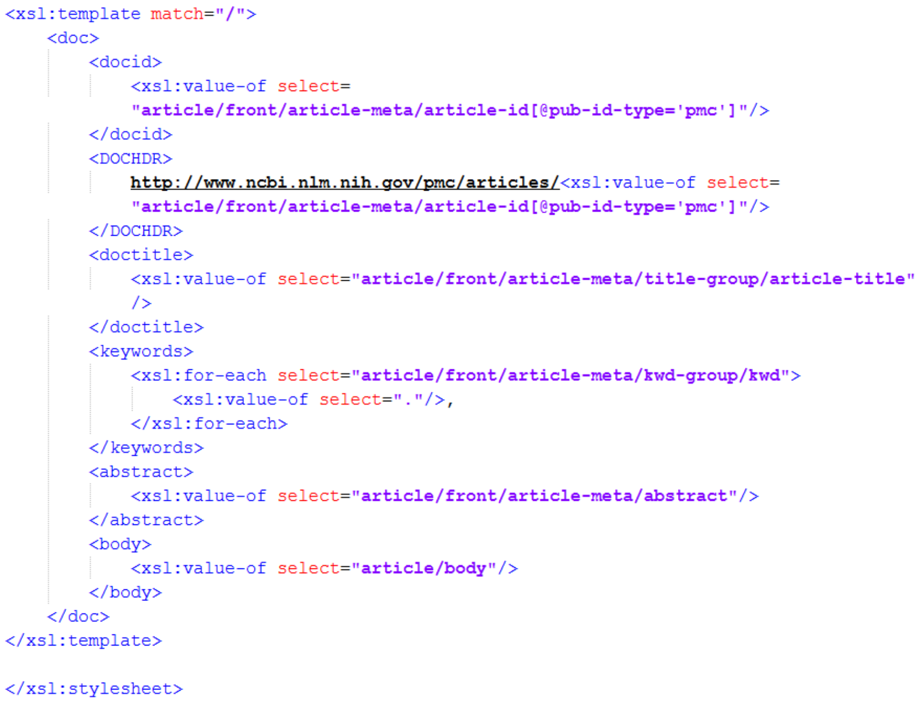
One can draw experience on effect of using lexicons from other semantic task areas.

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. However, situation is more complex with ontologies. A sophisticated analysis of Mesh-based expansion was given by [Lu et al., 2009]. One can either include Mesh terms (no automatic MeSH explosion) or include MeSH term and the more specific terms beneath it in the MeSH hierarchy, spanning several ontology levels. [Lu et al., 2009] concluded that MeSH automatic explosion feature improves results for some types of MeSH usage.

3 Retrieval methodology setup

3.1 Terrier

We are using the Terrier v4.1 [Terrier IR Platform] system to index provided set of documents and to retrieve results for a given set of queries. Terrier system provides vast amount of retrieval use cases including, but not limited to indexing a set of document and retrieving information from the set of documents. The Terrier system requires a set of uniformed XML documents to index. Format provided by PUB-MED is an extended format, which includes a lot of metadata, which is not required in neither retrieval nor indexing process. We are using a simple XSLT transformation to convert the provided data into a simpler, flat XML. The XSLT transformation selects an abstract, body, keywords and the id of the article. It also generates a link to the online version of the article. We put this information into a new xml file for each document. We omit articles without body and abstract. The transformation is presented on Figure 3.1.



```
<xsl:template match="/">
    <doc>
        <docid>
            <xsl:value-of select=
            "article/front/article-meta/article-id[@pub-id-type='pmc']"/>
        </docid>
        <DOCHDR>
            http://www.ncbi.nlm.nih.gov/pmc/articles/<xsl:value-of select=
            "article/front/article-meta/article-id[@pub-id-type='pmc']"/>
        </DOCHDR>
        <doctitle>
            <xsl:value-of select="article/front/article-meta/title-group/article-title"
            />
        </doctitle>
        <keywords>
            <xsl:for-each select="article/front/article-meta/kwd-group/kwd">
                <xsl:value-of select="."/>,
            </xsl:for-each>
        </keywords>
        <abstract>
            <xsl:value-of select="article/front/article-meta/abstract"/>
        </abstract>
        <body>
            <xsl:value-of select="article/body"/>
        </body>
    </doc>
</xsl:template>

</xsl:stylesheet>
```

Figure 3.1. XSLT transformation for converting the original article into a simpler version, which is used for indexing and eventually retrieval.

We are using a basic bash script to create and index provided with the Terrier software. We set up the indexing within the properties file. At this time we also set up TREC query tags for the expansions. We are using 3 document tags – DOC, DOCID and DOCHDR and a total of 12 distinct query tags – 3 basic tags (description, summary and note) and 9 additional tags for different types of expansions (relational knowledge expansion, fixed expansion and linguistic expansions). We use Terrier for the retrieval process. We use a set of configurations, various options for retrieval are described in further sections of this paper.

3.2 Query expansion details

We are using three specific expansion methods. The first method expands the query with a fixed vocabulary, which varies depending on the type of the query. Table 3.1

provides the vocabulary sets for each of the types. We call this method the Fixed Expansion.

Table 3.1. Vocabulary sets for various types of queries

Type	Vocabulary
Diagnosis	diagnose, analyse, examine analysis, conclusion, examination
Test	check, testing, checking
Treatment	prescribe, cure, hospitalization, medication, medicine, perform a surgery, operation, prescription, remedy, surgery, therapy, diet, therapeutic

The second type of expansions involves relational knowledge provided by the MeSH ontology. For this type, we run the script, which matches the terms within the text with classes with identical descriptions as the text from the query within the ontology. We expand the query with the vocabulary, which is a parallel description of the same class as well as with the descriptions of taxonomical children of that class down 3 levels. We call this method the RK Expansion.

Third type of expansions involves vector space models, such as word2vec language model [Mikolov et al., 2013a]). For this type of expansion we look for the three closest (in a cosine sense of distance) terms to the terms recognized and provided by the MeSH ontology. We expand the query with corresponding results. We have provided runs with various combinations of those three types. We call this method Linguistic Expansion.

4 Results

In this section we will go through the outcome of every retrieval setup implemented by our group and applied to the competition data sets. We will compare our results to median and best of the CDS submissions. Finally, we will discuss the best application for each setup. For the evaluation we will use NDCG and infAP measures. Importance and derivation of those inferred measures is described in [Yilmaz et al., 2008]. All of the presented evaluation results have been provided by TREC organizers. We have submitted 5 runs. Summary for the runs we provided is presented in Table 4.1

Table 4.1 Summary for the provided runs.

Run ID	Task	Method
DDPHBo1CM	Descriptions	DPH + Bo1 + Fixed Expansion + RK Expansion
DDPHBo1MWRe	Descriptions	DPH + Bo1 + Fixed Expansion + RK Expansion + Linguistic Expansion
SDPHBo1NE	Summaries	DPH + Bo1
NDPHBo1C	Notes	DPH + Bo1 + Fixed Expansion
NDPHBo1CM	Notes	DPH + Bo1 + Fixed Expansion + RK Expansion

4.1 Description task

In the following we present PUT results for the Description task.

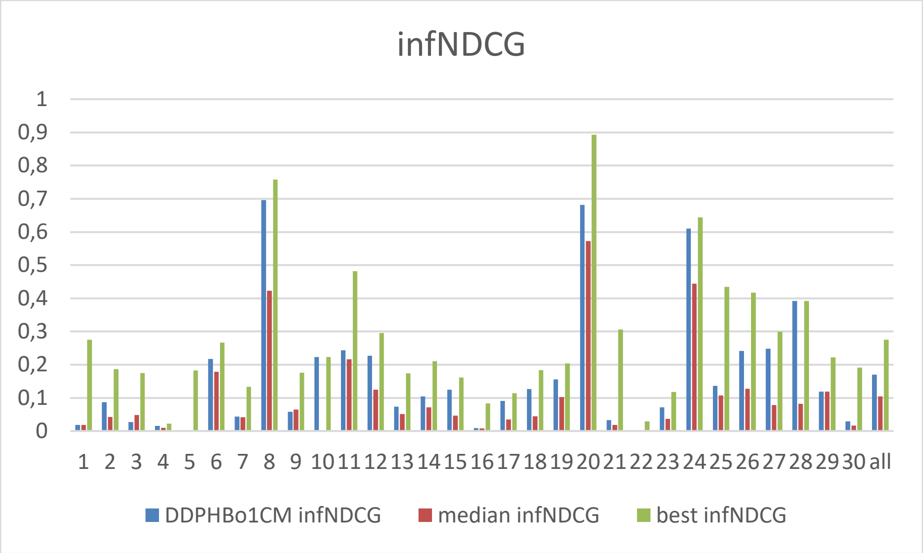


Figure 4.1 Distribution of infNDCG over the description task for the DDPHBo1CM setup (blue series) compared to median infNDCG in the description task (red series) and best infNDCG for this task(green series).

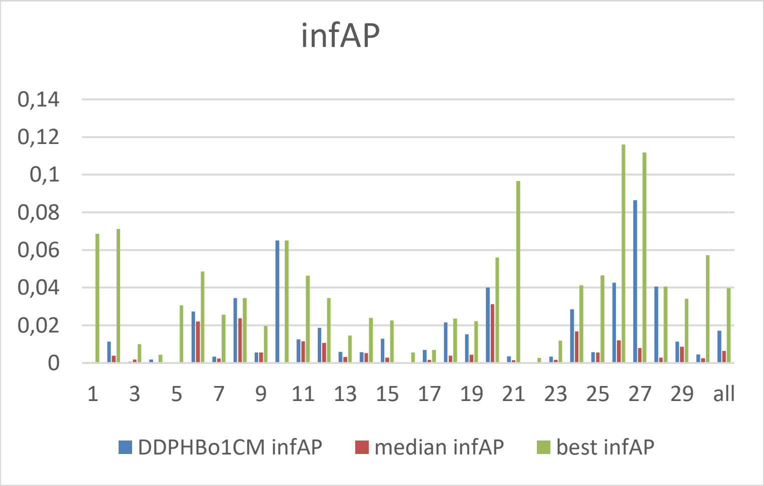


Figure 4.2 Distribution of infAP over the description task for the DDPHBo1CM setup (blue series) compared to median infAP in the description task (red series) and best infAP for this task(green series).

One of the competition tasks was involved with medium length, natural language description of the patient records. We have provided two distinct description run results. The first setup consists of basic Bo1 metric with pseudo relevance feedback. Additionally, this setup expands query with Fixed Extensions. We refer to the method as having id DDPHBo1CM. This basic method gives decent results. Figure 4.1. provides infNDCG measure across the topics with respect to this method, Figure 4.2. presents the same data for the infAP measure. We can observe that, for topics 8, 10, 17 and 28 this method was one of the best performing methods in the entire competition. The reason why there is so much difference between retrieval for specific topics is shown in Table 4.2. It provides extensive information about the texts this setup performs well together with a couple of cases for which the system performed badly. We can observe that topics, on which the setup performed well were descriptive in a common language sense. On the other hand the descriptions loaded with vast amount of professional, biomedical abbreviations cause this setup to perform poorly.

Table 4.2. Descriptions of topics selected for the qualitative analysis of the basic description case

Topic	Description
17	This is a 76-year-old female with personal history of diastolic congestive heart failure, atrial fibrillation on Coumadin, presenting with low hematocrit and shortness of breath. Her hematocrit dropped from 28 to 16.9 over the past 6 weeks with progressive shortness of breath, worse with exertion over the past two weeks. She reports orthopnea. She denies fevers, chills, chest pain, palpitaitons, cough, abdominal pain, constipation or diahrrea, melena, blood in her stool, dysuria or rash. Her electrocardiogram present no significant change from previous. Her Guaiac was reported as being positive.
28	This 84-year-old man with a history of coronary artery disease presents with 2 days of melena and black colored emesis. Stools becoming less dark, but he had increased lethargy and presented to the emergency department today. Initial systolic blood pressure recorded in the 60s, but all in 110-120s after that. In the ED, he had gastric lavage with coffee ground emesis that cleared with 600 cc of

	flushing. During the lavage he had chest pressure with mild ST depression V3-V5 that resolved spontaneously. Patient is on ASPIRIN 81 mg Tablet by mouth daily.

Table 4.2. (continuation) Descriptions of topics selected for the qualitative analysis of the basic description case

3	A 75F with a PMHx significant for severe PVD, CAD, DM, and CKD presented after being found down unresponsive at home. She was found to be hypoglycemic to 29 with hypotension and bradycardia. Her hypotension and confusion improved with hydration. She had a positive UA which eventually grew klebsiella. She had temp 96.3, respiratory rate 22, BP 102/26, a leukocytosis to 18 and a creatinine of 6 (baseline 2). Pt has blood cultures positive for group A streptococcus. On the day of transfer her blood pressure dropped to the 60s. She was anuric throughout the day. She received 80mg IV solumedrol this morning in the setting of low BPs and rare eos in urine. On arrival to the MICU pt was awake but drowsy. On ROS, pt denies pain, lightheadedness, headache, neck pain, sore throat, recent illness or sick contacts, cough, shortness of breath, chest discomfort, heartburn, abd pain, n/v, diarrhea, constipation, dysuria. Is a poor historian regarding how long she has had a rash on her legs.
8	A G2P0010 26 yo F, now estimated to 10 weeks pregnant, with 4yr hx of IDDM. Last menstrual period is not known but was sometime three months ago. Five days ago, the patient began feeling achy and congested. She had received a flu shot about 1 week prior. She continued to feel poorly and developed hyperemesis. She was seen in the ED (but not admitted), where she was given IVF, Reglan and Tylenol and she was found to have a positive pregnancy test. Today, she returned to the ED with worsening of symptoms. She was admitted to the OB service and given IVF and Reglan. Of note, her labwork demonstrates a blood glucose of 160, bicarbonate of 11, beta-hCG of 3373 and ketones in her urine. Her family noted that she was breathing rapidly and was quite somnolent. She appears to be in respiratory distress.

Our second run for this task is similar to the first one, however we have provided additional information in a form of RK Expansions and Linguistic Expansions. Let us refer to this method as to an extended method. We observe that multiple extensions frequently lower the quality of result. However, in multiple topics, extended queries gave better results, to achieve an overall lower quality of the answer. Nonetheless there are certain topics, for which the extended method was the best performing one. We present the comparison between the extended approach and the overall results in Figure 4.3. The differences between the basic and extended setup for description task are presented in Figure 4.4. We can see that the extended version performs similarly to the non-extended one. Overall the results are worse, however there are few topics, where the extended approach gives better results. Among others, topic 8, which is reported in Table 4.2 gives better results with the extended setup.

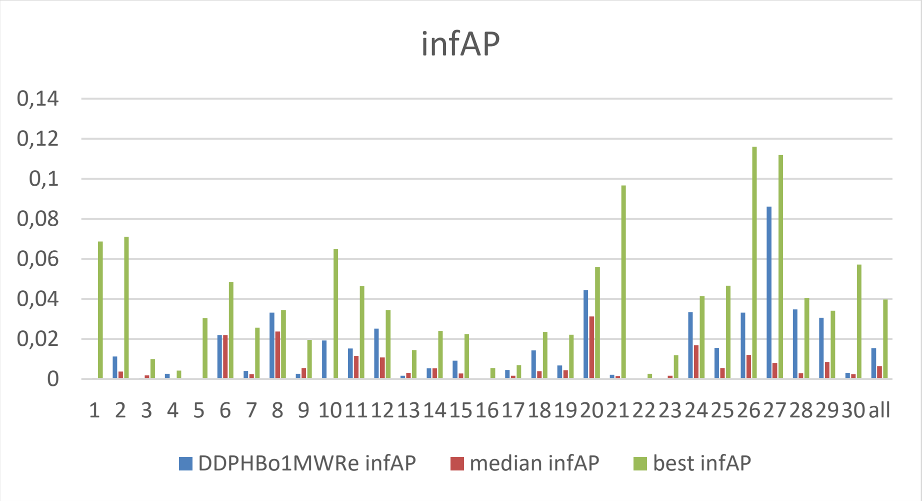


Figure 4.3 Distribution of infAP over the description task for the DDPHBo1MWRe setup (blue series) compared to median infAP in the Description task (red series) and best infAP for this task (green series).

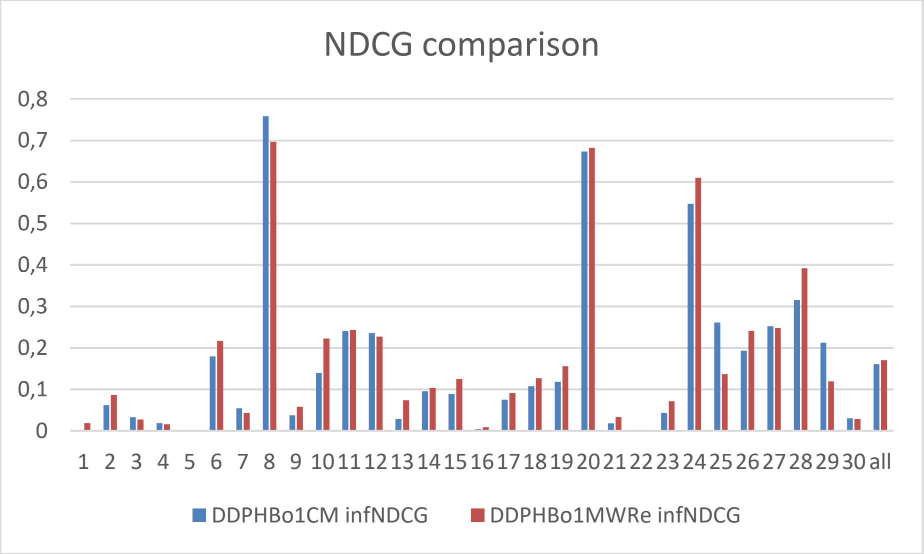


Figure 4.4. Comparison of inferred NDCG between both submitted Description task runs.

4.2 Summary task

The summary task in the competition was in the form of short texts in natural language as input to queries. These pieces of text, as well as the task itself are referred to as Summaries. For this task we have provided a run without any additional expansions. We use a DPH Bo1 method with Pseudo Relevance Feedback referred to as ID SDPHBo1NE. The inferred NDCG and inferred AP summaries are presented in Figure 4.5. and Figure 4.6. We can observe that this setup performs well across most of the topics, however it fails to come close to the best performing setups. The average measures for this setup are better than the median and results are consistent.

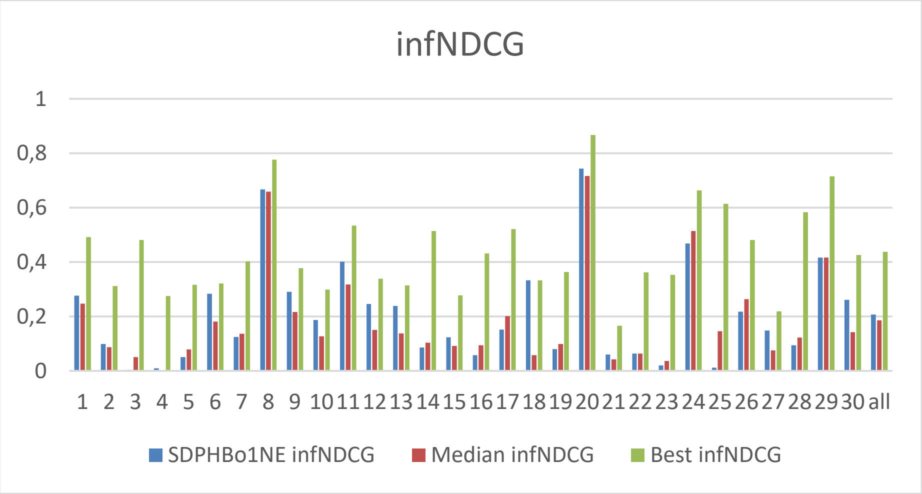


Figure 4.5 Distribution of infNDCG over the summaries task for the SDPHBo1NE setup (blue series) compared to median infNDCG in the summary task (red series) and best infNDCG for this task(green series).

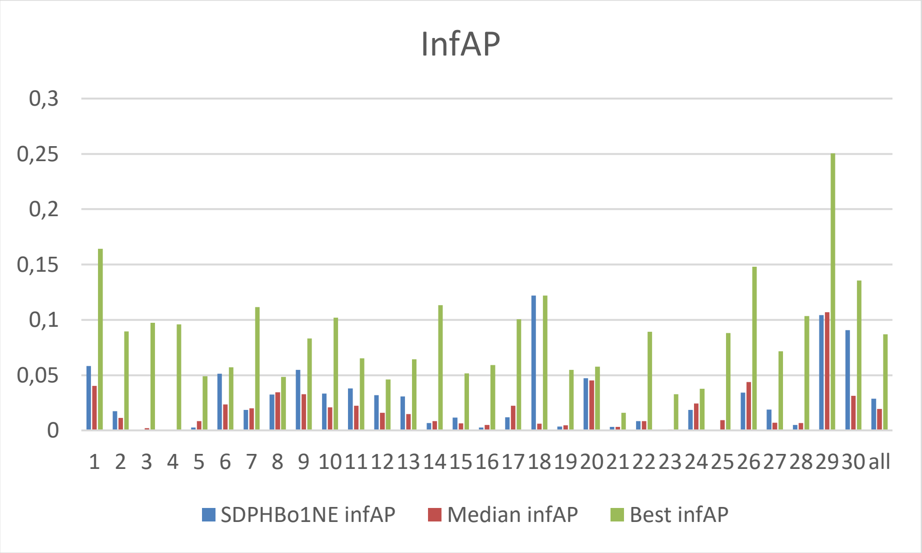


Figure 4.6. Distribution of infAP over the summaries task for the SDPHBo1NE setup (blue series) compared to median infAP in the summary task (red series) and best infAP for this task(green series).

4.3 Notes task

The Notes task of this year competition was to match the queries, which are similar to long, very descriptive texts. We have provided two runs for this task. The first one consists of a DPH Bo1 method with Pseudo Relevance Feedback and queries extended with Fixed Expansions. We refer to this method with NDPHBo1C identifier. Evaluation of inferred measures is presented in Figure 4.7. and Figure 4.8. We can see, that this setup provides good consistent results, however it fails to achieve the best results across. For the second run we have used the same setup with addition of the RK expansions. We refer to this method with identifier NDPHBo1CM identifier. Results for the second run are presented on Figure 4.9. and Figure 4.10. We can see that the expanded setup performs much better than the basic one. Evaluation measures are generally higher than the basic ones, however this setup achieves inferred measures close to the best performing methods.

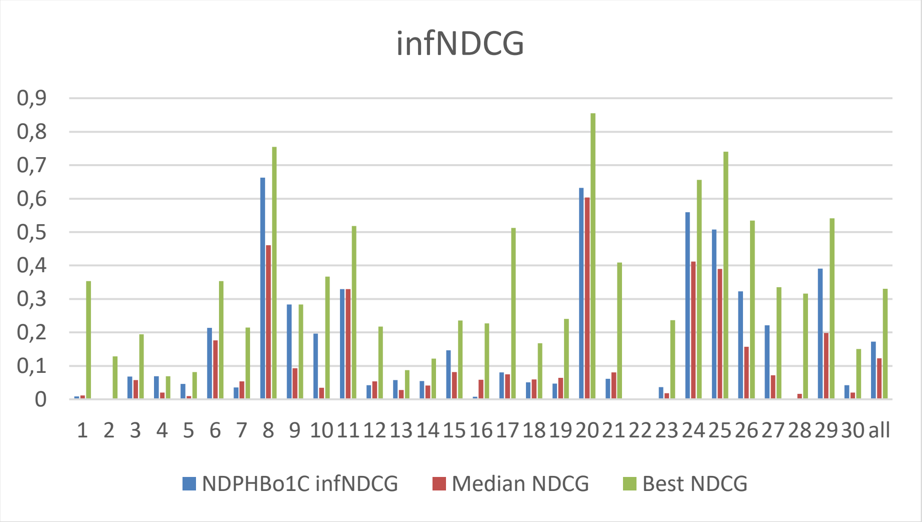


Figure 4.7. Distribution of infNDCG over the summaries task for the NDPHBo1C setup (blue series) compared to median infNDCG in the description task (red series) and best infNDCG for this task(green series).

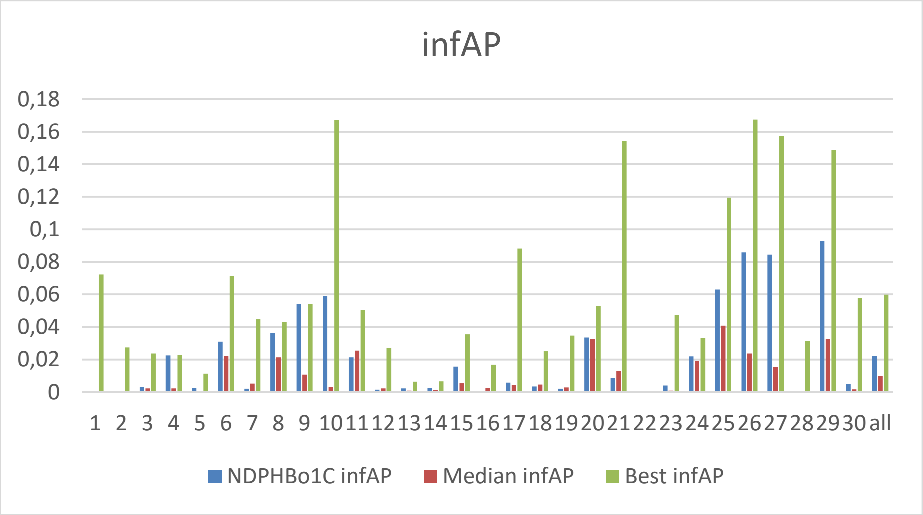


Figure 4.8. Distribution of infAP over the summaries task for the NDPHBo1C setup (blue series) compared to median infAP in the description task (red series) and best infAP for this task(green series).

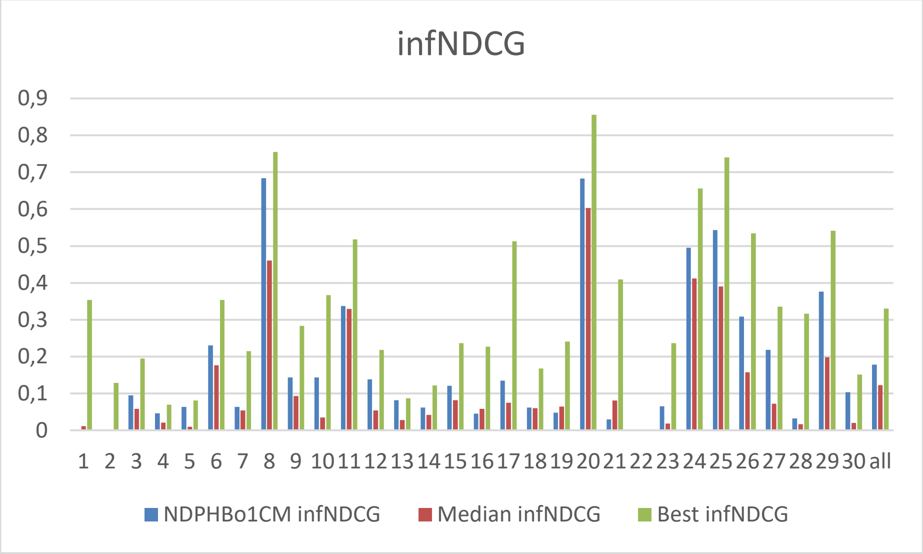


Figure 4.9. infNDCG distribution for the extended version notes task – blue series present data achieved for our system; red series present median data across the competition; green series present the best performance for a given topic.

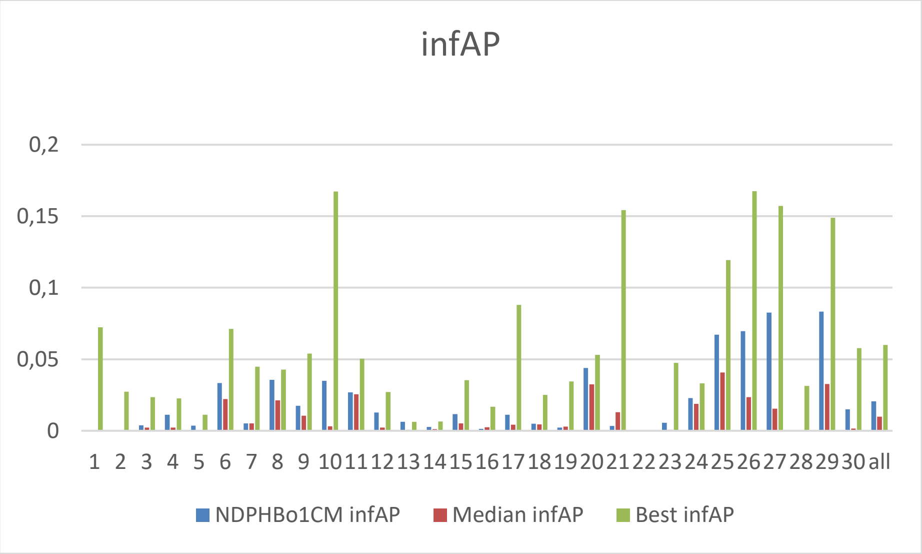


Figure 4.10. infAP distribution for the extended version of notes task – blue series present data achieved for our system; red series present median data across the competition; green series present the best performance for a given topic.

5 Conclusions and future work

In all measures our results are approximately 0,02 above the median, which is mostly due to the Terrier DPH Bo1 method chosen as a baseline. We also used Mesh query expansion, word2vec query expansion, and the combination of these two options. The Mesh query expansion improves the results roughly by 0,008, and the word2vec causes the results worse with regard to the baseline.

It is to be stressed that in [Mikolov et al., 2013a] the pure word2vec method was presented as better than it actually by choosing an easy type of a corpus such as countries and capitals. Much better results are obtained when sense disambiguation and hubness reduction is applied to the vector space [Faruqui et.al., 2015], [Jauhar et.al., 2015], [Dutkiewicz, Jedrzejek, 2016]. We intend to use these word embedding improvement to search in future. Also one could uses word embedding by calculating similarities between input and target documents [Mikolov et al., 2013b,], [Le, Mikolov, 2014].

We did not provide the source code for our method. In future we will adhere to Open Runs with TREC submissions that are backed by an open-source code repository such that someone can check out and execute the run as submitted to TREC.

This work was supported by the PUT DS grant no 04/45/DSPB/0149.

References

ACLWEB TOEFL, (2016), TOEFL Synonym Questions (State of the art)https://www.aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)

ACLWEB ESL (2016), ESL Synonym Questions (State of the art) https://www.aclweb.org/aclwiki/index.php?title=ESL_Synonym_Questions_(State_of_the_art)

Dutkiewicz J. , Jedrzejek C., (2016) Query Answering to IQ Test Questions Using Word Embedding with Retrofit and Hubness Reduction, CoRR abs (2016)

Frąckowiak M., Dutkiewicz J., Jędrzejek C., Retinger M. Werda P., (2016) Query Answering to IQ Test Questions Using Word Embedding, in Multimedia and Network Information Systems, Volume 506 of the series Advances in Intelligent Systems and Computing pp 283-294

Goodwin T., Harabagiu S. M, UTD at TREC 2014, (uery Expansion for Clinical Decision Support. In Ellen M. Voorhees, Angela Ellis: *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014.* National Institute of Standards and Technology (NIST) 2014 TREC 2014

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

Jauhar S. K., Dyer C., Hovy E. H., (2015) Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models. HLT-NAACL pp.683-693

Le Q. V., Mikolov T., (2014) Distributed Representations of Sentences and Documents. ICML 2014: 1188-1196

Lin J., Crane M. , Trotman A., Callan J., Chattopadhyaya I., Foley J., Ingersoll G., Macdonald C., and Vigna S., (2016), Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge, in LNCS 9626, FerroN. et al. (Eds.), Springer 2016, pp. 408–420

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, Corrado, G.S., and Dean, J. (2013a), Distributed Representations of Words and Phrases and their Compositionality. NIPS, pp. 3111-3119

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

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.

Terrier IR Platform *www.terrier.org* 26 Oct 2016

Yilmaz E., KanoulasE., and. Aslam J. A. (2008). 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* (SIGIR '08). ACM, New York, NY, USA, 603-610.

Wang, Y., & Fang, H. (2014). Explore the query expansion methods for concept based representation. In *Proceedings of the 2014 Text Retrieval Conference*.