DUTH at TREC 2015 Clinical Decision Support Track*

George Drosatos[†1], Stefanos Roumeliotis[1], Avi Arampatzis[2] and Eleni Kaldoudi[1]

[1]School of Medicine, Democritus University of Thrace, Alexandroupoli, Greece
gdrosato@ee.duth.gr, st_roumeliotis@hotmail.com, kaldoudi@med.duth.gr
[2]Electr. & Comp. Eng. Dept., Democritus University of Thrace, Xanthi, Greece
avi@ee.duth.gr

Abstract

In this report we give an overview of our participation in the TREC 2015 Clinical Decision Support Track. We present two approaches for pre-processing and indexing of the open-access PubMed articles, and four methods for query construction which are applied to the previous two approaches. Regarding pre-processing, our main assumption is that only particular medical study designs are appropriate for each type of clinical question and we filter the number of articles in each clinical question type. Regarding query construction, our main idea is to detect the medical concepts in the medical cases and to expand them with terms of semantic controlled vocabularies (such as UMLS). The track evaluation shows that our approaches provide comparable results with the other participants' approaches without to conclude on safe findings.

1 Introduction

TREC 2015 is the second year that the Clinical Decision Support Track is running. The track's goal is to investigate techniques for linking medical cases to biomedical articles relevant for patient care. In other words, the track's challenge is to retrieve for a given case report full-text biomedical articles that answer questions related to several types of clinical information needs. This year there were two tasks (A and B) in contrast with the 2014 track which only had one task. Each task consists of 30 topics that are equally separated into three types: diagnosis, test, and treatment. Each topic, apart from the type, consists of a description and a summary that represent a medical case (or synonymously "case report"). Task A was exactly the same with 2014's task, and task B additionally provided a diagnosis field for the treatment and test topics. Based on the type of a topic, the clinical questions that we had to answer with relevant articles (a ranked list of 1000 documents) and accordingly with the topic's type are: "What is the patient's diagnosis", "What tests should the patient receive?" and "How should the patient be treated?".

Regarding the corpus/collection of biomedical articles, the track provides a subset of articles from PubMed[1]. This subset consists of open-access full-text biomedical articles that are provided

*In: Proceedings of the Text REtrieval Conference (TREC), NIST, 2015.
[†]corresponding author
[1]`http://www.ncbi.nlm.nih.gov/pubmed`

by the PubMed Central[2] (PMC). A snapshot of this subset was obtained on January 21, 2014 and contains a total of 733,138 articles. The same snapshot was used in both tracks of 2014 and 2015.

The remainder of this report is organized as follows. Section 2 describes our methodology for pre-processing and indexing the articles. Our proposed query construction methods are presented in Section 3. Our submitted runs and the official results of TREC Clinical Decision Support Track are described in Section 4. Finally, Section 5 draws our conclusions.

2 Preprocessing and Indexing

In this section, we discuss in detail the pre-processing and indexing methods we employed in order to create our indices of articles for our experiments. Our goal was to build two different types of indices for our retrieval techniques that are presented in the next section. These two type of indices are:

(a) **Dedicated indices per type of clinical question based on study designs:** This idea to build indices per type of clinical question is based on the assumption that particular types of medical studies are appropriate to answer the different types of clinical questions. For this reason, we tried to classify the biomedical articles into particular study designs and to link these classes to the different types of clinical questions.

(b) **General index for all clinical questions:** This type of index consists of all the biomedical articles and it is used in order to be possible to compare our results with the type (a) of indices.

Figure 1 shows an overview of our preprocessing and indexing procedures in order to build these two types of indices. As shown in Figure 1, in both types of indices we exclude all the articles that did not have body (full-text), e.g. only abstracts or some articles with copyrighted contents. Thus, the remaining documents for indexing are in total 642,250 articles.

2.1 Indices of Type (a)

In order to classify the articles, we introduce the "EDDA Study Designs and Publications" (EDDA) ontology[3] that is based on research described in [2]. This ontology gives terminology of study designs and publication types, and in our case we decided to only focus on the study designs that represent the majority of different categories of medical studies. The defined categories of studies in version 1.3 of the ontology are 172 categories (including subcategories) that are enriched with synonym terms from MeSH, NCI Thesaurus (NCIT), and Emtree, the controlled vocabularies for MEDLINE, the National Cancer Institute, and Embase, respectively.

The list of these studies was given to a medical expert who was requested to classify each study design as "good" or "bad" for the three types of clinical questions (diagnosis, test and treatment). Subsequently, the results of this process were used to generate "good" and "bad" queries per type of clinical questions. The good query consists of study designs, as phrases or terms, expanded with their synonyms where the doctor gave positive answer for a particular clinical question type and the bad query has the opposite study designs. An example of a good and a bad query for diagnosis using the Indri Query Language [8] is:

[2] http://www.ncbi.nlm.nih.gov/pmc

[3] http://purl.bioontology.org/ontology/EDDA

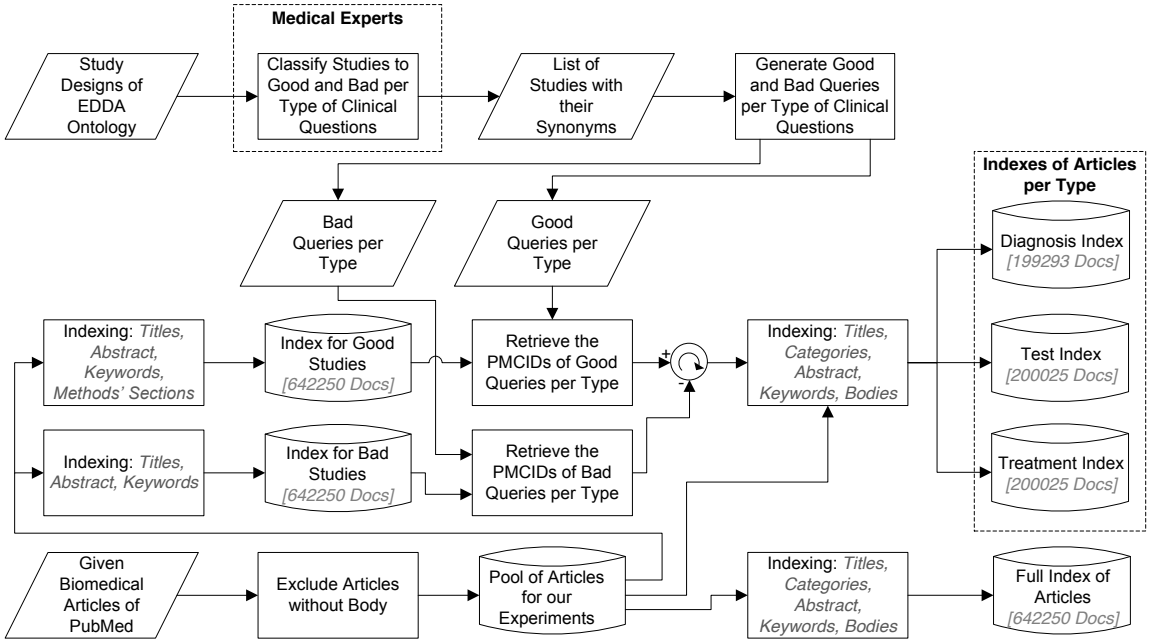

Figure 1: A system flowchart of the preprocessing and indexing procedures of articles.

Good Query:

```
#combine(<#2(case series) #2(case series trial)> ··· <survey>)
```

Bad Query:

```
#combine(<#1(animal study) #1(animal experiment)> ··· <#1(in vivo study)>)
```

In the previous queries, the expression $< \cdots >$ defines a list of synonyms terms, the expression #1(*aaa bbbb*) matches "aaa bbbb" as an exact phrase and the expression #2(*aaa bbbb*) matches "aaa $*$ bbbb" (where $*$ is any word or null). In order to avoid some fault detections of bad documents, we use stricter matching expression (#1) for phrases in bad queries than the expression (#2) in good queries.

Then, the generated queries are used to retrieve the PubMed Central IDs (PMCIDs) of all the documents that match with the queries. For the "bad" queries, we use an index that consists of titles, abstracts and keywords, and for the "good" queries we additionally use the section of methods. The extraction of methods' section is achieved in some cases because this section is explicitly defined as 'methods' in the XML files of articles and in some other cases because we detect the term 'method' in the title of sections. We use this variation in the indices for the "good" and "bad" queries in order to avoid some fault detections of bad documents from the methods' sections because in the methodology of an article is possible to describe that the presented study is based on other previous, for example, animal studies.

We used the Indri v5.5 [8], as search engine and indexing builder, the indexing of articles is performed using the default settings of this version, except that we enable the Krovetz stemmer [6], and the ranking of retrieving results is performed using the default (Language Model (LM) [8])

retrieval model of Indri. The resulted PMCIDs of the good and bad queries per clinical question type are combined (excluding the bad PMCIDs from the good) and using the articles that contain full-text, we build three indices for each clinical question type. These indices consist of the following XML fields: titles, publication categories, abstracts, keywords and bodies (full-text).

2.2 Index of Type (b)

In this type of index, we do not follow any special technique in the building process and for indexing we use all the articles that have body (full-text) using the default settings of Indri v5.5 [8], except that we enable the Krovetz stemmer [6]. The XML fields that are used in this case is the same with the indices of type (a), i.e. title group, publication categories, abstract, keywords and body (full-text). The purpose of this index is to used as a reference point in our experiments and to detect the differences from the usage of indices in type (a).

3 Query Construction Methods

In this section, we present the four query construction methods that are applied to the two types of indices (Section 2) in order to provide three runs for each task (A and B) of the TREC 2015 clinical decision support track. For retrieval, we use as a search engine the Indri v5.5 [8], and the ranking of retrieving results is performed using the default (Language Model (LM)) retrieval model. The proposed four query construction methods are:

(i) Manual query construction from medical experts

(ii) Automatic query construction using all summary terms of each topic

(iii) Automatic query construction using the semantically expanded summary terms of each topic that are previously mapped to particular medical concepts

(iv) Automatic query construction using the method (iii) and additionally the semantically expanded terms of clinical question types and diagnosis field (only in task B)

In the following we describe in detail the proposed four query construction methods.

Method (i): In this query construction method, we recruited a medical expert to provide us with the queries he would issue in PubMed's search engine in order to retrieve relevant articles for each medical case and clinical question. The majority of these queries were practically the diagnoses of topics. The main goal of this method is to understand how medical experts search in PubMed to find helpful literature. This manual query construction method consists our manual run in the task A of the competition.

Method (ii): In this method, we use as query in each topic all the terms in an ORed fashion (`#combine( ··· )`) of the topic's summary field by removing all punctuation and special characters. The produced queries of this method were used to retrieve a baseline set of results. However, the track results showed, as shown in Section 4, that this simple method works better than the other more complicated methods.

Method (iii): An overview of this method is shown in Figure 2 and follows an approach somehow similar to [4] regarding the utilized semantic tools. The main goal of this method is to detect the

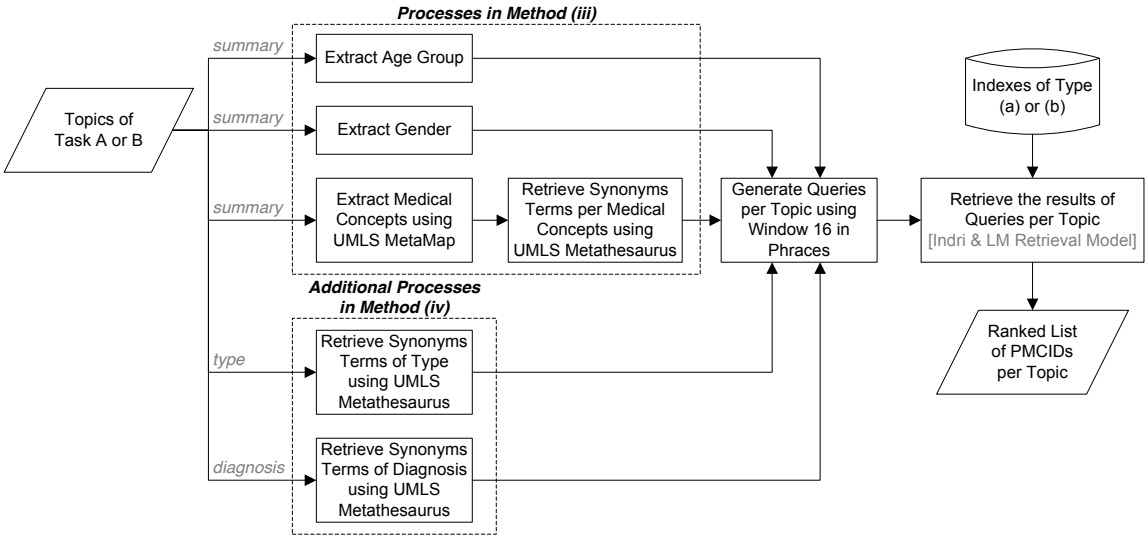

Figure 2: An overview of the proposed query construction methods (iii) and (iv).

medical concepts of the topic's summary and to expand them with terms from semantic control vocabularies. The processes that we followed in order to perform the query construction are:

- *Age group extraction:* In this process, we detect keywords or phrases, such as "XX-year-old", "XX yo", "young", and "child", in the summary field of topics, and assign them to particular age groups: 'child' (age $\leq$ 16), 'young adult' (17 $\leq$ age $\leq$ 45), 'middle age adult' (46 $\leq$ age $\leq$ 65) and 'old adult' (age $\geq$ 66).
- *Gender extraction:* Accordingly, with this process we detect the subject's gender in the summary of topics. Thus, if we detect the term 'man' or 'male', we will keep in the query the synonyms terms `<man male>` and so on.
- *Medical concepts extraction:* In order to detect automatically the medical concepts in the topics' summary, we utilize the MetaMap [1] application of UMLS [3] through its Java API[4]. The parameters of MetaMap application were
  - Knowledge source: *2014AB (1415)*
  - Data version: *USAbase*
  - Data model: *Strict Model (-A)*
  - Restrict to semantic types: *Body Location or Region*, *Disease or Syndrome*, *Eukaryote*, *Finding*, *Injury or Poisoning*, *Mental or Behavioral Dysfunction*, *Pathologic Function*, *Pharmacologic Substance*, *Sign or Symptom*, *Therapeutic or Preventive Procedure* and *Virus*.

  The results of this process were a list of UMLS Concept Unique Identifiers (CUIs) per matched medical concepts.

[4] `http://metamap.nlm.nih.gov/JavaApi.shtml`

- *Synonyms terms expansion of medical concepts:* The mapped medical concepts (list of CUIs) from the previous process are used as input in the UMLS Metathesaurus application [3] through its Java API[5] in order to retrieve synonyms terms. The used parameters in this application were:
  - UMLS release: *2014AB*
  - Included language of terms: *English (ENG)*
  - Include obsolete terms: *false*
  - Included term types[6]: *Designated synonym (SY)*, *Designated preferred name (PT)*, *Metathesaurus preferred name (PN)*, *Metathesaurus-supplied form of British synonym (MTH_SYGB)*, *MTH Designated synonym (MTH_SY)*, *Root synonym (RSY)*, *Term that is related to (RT)*, *Synonymous"short" forms (SS)*, *British synonym (SYGB)* and *Unique synonym (USY)*.

In all previous processes, the final detected terms, phrases or synonyms are passed through the Krovetz stemmer [6] in order to avoid duplicated terms or phrases in the generated query. Thus, the final generated query consists of age group, gender and only the mapped medical concepts with their synonyms. A query example of this method using the Indri Query Language [8] has the following form:

```
#combine(young adult <man male> <#16(coffee ground emesi) ··· > ··· )
```

In this query and in case where there are phrases, we use a maximum ordered window of 16 words instead of the exact phrase so as to relax the search condition of phrase medical concepts. This window of 16 words, as shown in literature [9], performs better results and ensuring some relatedness between terms.

Method (iv): In comparison with method (iii), this method additionally exploits the topic's diagnosis using semantically expanded terms in order to enrich the retrieval results. Apart from the processes in the method (iii) and as shown in Figure 2, we employ two additional processes in the query generation:

- *Adding expanded clinical question types:* For the three different types, we include in the query the following synonyms terms based on the UMLS Metathesaurus:
  - Diagnosis (CUI: C0011900): `<diagnosis diagnose diagnosing>`
  - Test (CUI: C0039593): `<test testing>`
  - Treatment (CUI: C1522326): `<treatment treat treating therapy management>`
- *Synonyms terms expansion of diagnosis:* Accordingly with the processes *'Medical concepts extraction'* and *'Synonyms terms expansion of medical concepts'* in method (iii), we map the diagnosis (available in test and treatment types of task B) with its CUI and expand it with synonyms terms.

The final generated query has the same form with the query in method (iii) including additionally the resulted terms from the previous two processes.

[5] `https://uts.nlm.nih.gov/home.html#apidocumentation`
[6] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html`

Table 1: The submitted runs in comparison with the index type and the query construction method.

Task	Runs	Run Type		Index Type		Query Construction Method			
		Man	Auto	(a)	(b)	(i)	(ii)	(iii)	(iv)
Task A	DuthStef	✓			✓	✓			
	DuthMmMt16s		✓	✓				✓	
	DuthMmMt16f		✓		✓			✓	
Task B	DuthBaseS		✓	✓			✓		
	DuthBaseF		✓		✓		✓		
	DuthMmMtB16f		✓		✓				✓

4 Runs and Results

In this section, we discuss the features as well as the official evaluation of the six runs we submitted to the TREC 2015 Clinical Decision Support Track.

4.1 Submitted Runs

In the TREC 2015 Clinical Decision Support Track we submitted in total six runs, as shown in Table 1, based on two different index types (Section 2) and different query construction method (Section 3). In the submitted runs we followed the guidelines given in the web-page of the track [10].

4.2 Results

The evaluation results according to the infAP, infNDCG, R-prec, and P@10 measures over all the topics of our six runs are reported in Table 2. Additionally, Figures 3 and 4 show as box-and-whisker plots the distribution of infNDCG and P@10 scores, respectively, across all the topics for our runs. Runs plotted with blue shading are manual runs, with white shading are automatic runs that utilize the diagnosis field of topics and with gray shading are automatic runs that are comparable among them. The definitions of these measures are:

- *Inferred Average Precision (infAP):* This measure approximates the average precision even when the relevance judgments are incomplete [11, 12].
- *Inferred Normalized Discounted Cumulative Gain (infNDCG):* This measure approximates the usefulness, or gain, of a document based on its position in the result list using the graded relevance scale of the assessors [5, 12].
- *R-Precision (R-prec):* R-precision is the precision after R documents have been retrieved, where R is the number of relevant documents for the topic.
- *Precision at Rank 10 (P@10):* The fraction of articles within the top-10 results where the doctor judges as relevant documents.

Table 2: Mean of results over all the topics for infAP, infNDCG, R-prec and P@10 measures.

Topic type	Runs	infAP	infNDCG	R-prec	P@10
Total	DuthStef	.0429	.2318	.1379	.3500
	DuthMmMt16s	.0265	.1620	.1071	.2900
	DuthMmMt16f	.0404	.1849	.1406	.3133
	DuthBaseS	.0305	.1723	.1218	.3500
	DuthBaseF	.0446	.2105	.1685	**.3933**
	DuthMmMtB16f	**.0475**	**.2394**	**.1784**	.3800
Diagnosis	DuthStef	**.0354**	**.2463**	.1298	.3100
	DuthMmMt16s	.0165	.1495	.0958	.3000
	DuthMmMt16f	.0201	.1567	.1217	.3200
	DuthBaseS	.0246	.1713	.1322	.3600
	DuthBaseF	.0340	.2102	**.1630**	**.4000**
	DuthMmMtB16f	.0233	.1715	.1326	.3200
Test	DuthStef	.0182	.1760	.0799	.2700
	DuthMmMt16s	.0172	.1500	.0948	.2300
	DuthMmMt16f	.0254	.1729	.1207	.2800
	DuthBaseS	.0186	.1594	.1014	.2800
	DuthBaseF	.0275	.1947	.1524	.3500
	DuthMmMtB16f	**.0458**	**.2758**	**.1909**	**.4100**
Treatment	DuthStef	.0752	**.2731**	.2043	**.4700**
	DuthMmMt16s	.0457	.1863	.1309	.3400
	DuthMmMt16f	**.0757**	.2251	.1795	.3400
	DuthBaseS	.0482	.1863	.1317	.4100
	DuthBaseF	.0722	.2267	.1900	.4300
	DuthMmMtB16f	.0732	.2709	**.2117**	.4100

According to Table 2, and Figures 3 and 4, `DuthMmMtB16f` yielded better results in all topics and tasks (A and B) than the other in 3 of 4 evaluation measures (infAP, infNDCG, and R-prec). However, this run is not fair to compare with the other five runs because is the only run that utilize the diagnosis field of test and treatment topics. The four runs, `DuthMmMt16s`, `DuthMmMt16f`, `DuthBaseS` and `DuthBaseF`, are comparable among them as automatic runs and `DuthBaseF` provided better results in all evaluation measures. From this comparison, we concluded that the simple query construction method (ii) yields better results than the semantic-based method (iii) and also the simple index of type (b) is better than the dedicated indices of type (a) based on study designs of articles. As regards the comparison with the manual run `DuthStef`, the automatic runs, `DuthBaseF` and `DuthMmMtB16f`, provided comparable results with this manual run and some measures (infAP, R-prec and P@10) have better results from this. Another interesting conclusion is that `DuthMmMtB16f` is little better than `DuthMmMt16f` for the clinical question of diagnosis and the only difference in the queries was the synonyms set `<diagnosis diagnose diagnosing>`. Overall, the evaluation results of our participation show that our runs (or some of that) are comparable with the runs of other participants (this year track and 2014 track [7]) and sometimes simpler solutions are better in result quality.

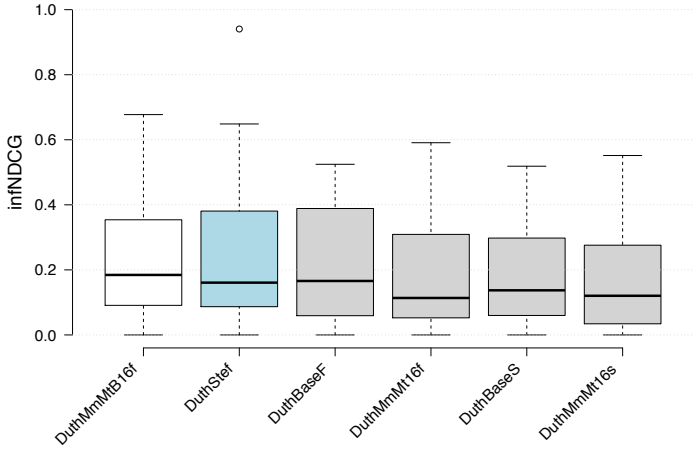

Figure 3: Distribution of infNDCG scores for our runs ordered by decreasing mean.

5 Conclusions

In this paper we described our participation in the TREC 2015 Clinical Decision Support Track. In the context of this work, we presented two approaches for the pre-processing and indexing of the open-access PubMed articles, and four methods for the query construction. As regards the pre-processing, our main assumption was that only particular medical study designs are appropriate for each type of clinical question and tried to reduce the number of articles in each clinical question type. Accordingly, as regards the query construction, our main idea was to detect the medical concepts in the medical reports/cases and to expand them with terms of semantic controlled vocabularies (such as UMLS). However, the evaluation results showed that sometimes the simplest solutions are the best.

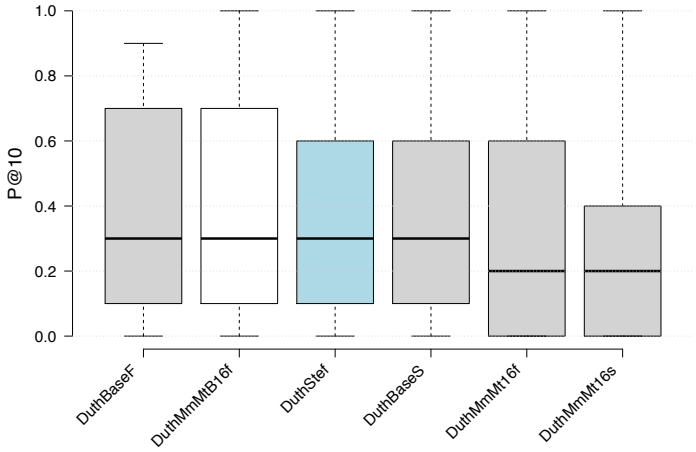

Figure 4: Distribution of P@10 scores for our runs ordered by decreasing mean.

Acknowledgements. The present work was partially funded by by the FP7-ICT project CARRE (Grant No. 611140), funded in part by the European Commission.

References

[1] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] T. Bekhuis, D. Demner-Fushman, and R. S. Crowley. Comparative effectiveness research designs: an analysis of terms and coverage in medical subject headings (mesh) and emtree. *Journal of the Medical Library Association: JMLA*, 101(2):92, 2013.

[3] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[4] J. I. Garcia-Gathrighta, F. Menga, and W. Hsua. Ucla at trec 2014 clinical decision support track: Exploring language models, query expansion, and boosting. In *Proceedings of 23rd Text Retrieval Conference (TREC)*. National Institute of Standards and Technology (NIST), 2014.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[6] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93)*, pages 191–202, New York, NY, USA, 1993. ACM.

[7] M. S. Simpson, E. Voorhees, and W. Hersh. Overview of the trec 2014 clinical decision support track. In *Proceedings of 23rd Text Retrieval Conference (TREC)*. National Institute of Standards and Technology (NIST), 2014.

[8] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6, 2005.

[9] E. Terra and C. L. Clarke. Frequency estimates for statistical word similarity measures. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pages 165–172. Association for Computational Linguistics, 2003.

[10] TREC Clinical Decision Support. Trec 2015 clinical decision support track guidelines, October 2015. `http://trec-cds.appspot.com/2015.html`.

[11] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111. ACM, 2006.

[12] E. Yilmaz, E. Kanoulas, and J. A. Aslam. 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*, pages 603–610. ACM, 2008.