A Hybrid Approach to Precision Medicine-related Biomedical Article Retrieval and Clinical Trial Matching

Yuan Ling[1], Sadid A. Hasan[1], Michele Filannino[2,4], Kevin P. Buchan[3],
Kahyun Lee[4], Joey Liu[1], William Boag[2], Di Jin[2], Özlem Uzuner[2,4],
Kathy Lee[1], Vivek Datla[1], Ashequl Qadir[1], and Oladimeji Farri[1]

[1]Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
`{firstname.lastname,kathy.lee_1,dimeji.farri}@philips.com`
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
`{wboag,jindi15}@mit.edu`
[3]State University of New York, Albany, NY, USA
`kbuchan@albany.edu`
[4]George Mason University, Fairfax, VA, USA
`{mfilanni,klee70,ouzuner}@gmu.edu`

Abstract

We describe our systems implemented for the Text Retrieval Conference (TREC 2017) Precision Medicine track. We submitted five runs for biomedical article retrieval and five runs for clinical trial matching. Our approaches combine strict rule matching with an ontology-based solution. Evaluation results demonstrate that our best run obtained the 2nd highest precision (P@5) score for the clinical trial matching task and was consistently ranked within top 5 teams in all evaluation metrics for the biomedical literature retrieval task.

1 Introduction

The Precision Medicine Initiative promoted by the Obama administration promised "to support doctors in delivering the right treatments, at the right time, every time to the right person" (Obama, 2015). In that spirit, the organizers of the TREC 2017 Precision Medicine track[1] challenged participants to build automatic retrieval systems capable of matching clinical scenarios with relevant treatments presented in the abstracts of biomedical scientific articles (Task 1), and relevant clinical trials that meet the patients' eligibility criteria (Task 2). Building such automated retrieval systems is essential as the process of manually searching biomedical articles and matching clinical trials requires significant medical

[1]http://www.trec-cds.org/2017.html

research and investigation based on an underlying clinical scenario.

The TREC 2017 Precision Medicine track focused on providing useful precision medicine-oriented information to clinicians related to treatments of cancer patients. In particular, the clinical scenario of a patient's disease was composed by cancer type, relevant genes or their variants, demographic information (age and gender), and other pertinent factors. We submitted five runs for biomedical article retrieval and five runs for clinical trial matching. Our approaches combined strict rule matching, textual similarity measures and ontologies. The official evaluation results demonstrated the effectiveness of our systems. In the next sections, we describe our systems, discuss experimental setup, and present evaluation results with analyses.

2 System Description

We build a hybrid system to retrieve and match treatment-related biomedical articles and clinical trials to eligible patients based on a strict rule matching approach and an ontology-based approach. We describe these approaches in the next subsections.

2.1 Strict Rule Matching-based Approach

Our first approach is based on strict rule matching principles. Figure 1 shows the basic architecture of our system, which includes 7 modules. The clinical trial indexing module and PubMed article indexing module index large volumes of clinical trials and PubMed articles into ElasticSearch. Clinical
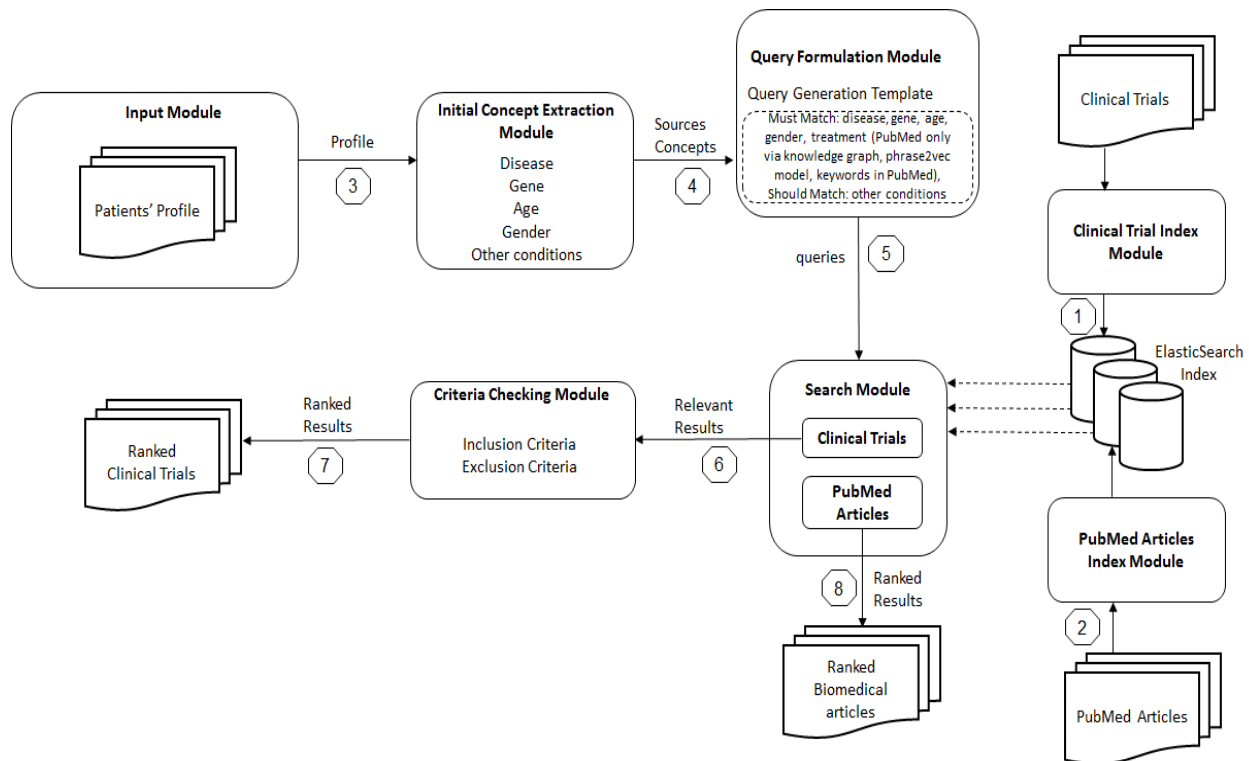
Figure 1: System architecture for strict rule matching-based approach

trial indexing considers fields such as *title*, *gender*, *age*, *criteria* etc. whereas PubMed article indexing considers the fields: *title* and *abstract*.

The input module takes the given topics (i.e. structured patient scenarios), along with external resources and strict extraction rules (designed to reformat the concepts from topics, i.e. *diseases*, *genes*, *age*, *gender*, and *other conditions*). The extracted concepts are fitted into the query formulation module such that relevant queries are generated for searching in ElasticSearch. The concepts for *diseases*, *genes*, *age*, *gender*, and *treatments* (for PubMed article search only) are set as *must match* for queries. The concepts for other conditions are set as *should match* for queries.

The search module for clinical trial matching returns a list of possibly relevant clinical trials (candidates ranked by ElasticSearch), which then goes through the criteria checking module. This module checks the inclusion and exclusion criteria section of the candidate trials, and excludes the clinical trials indicating any patient conditions in the exclusion criteria, then returns the final ranked clinical trials

for patients. Similarly, the PubMed search module returns the ranked PubMed articles using a strict-rule matching algorithm combined with additional matching criteria set using a list of relevant treatment concepts. These treatment concepts are inferred by providing disease concepts to a Wikipedia-based knowledge graph (Datla et al., 2017), and genes to a phrase2vec model (Hasan et al., 2015) trained on the given PubMed collection.

2.2 Ontology-based Approach

This approach at first filters out completely irrelevant biomedical articles (i.e. any non-cancer related article) by discarding the ones not related to neoplasms[2] (C.04* among the MeSH ontology codes). This allows the system to focus more on the algorithm rather than scalability. The system then enriches the information provided in the topics with synonyms for genes (using Entrez Gene Library[3]), diseases and other conditions (using MEDLINE[4]).
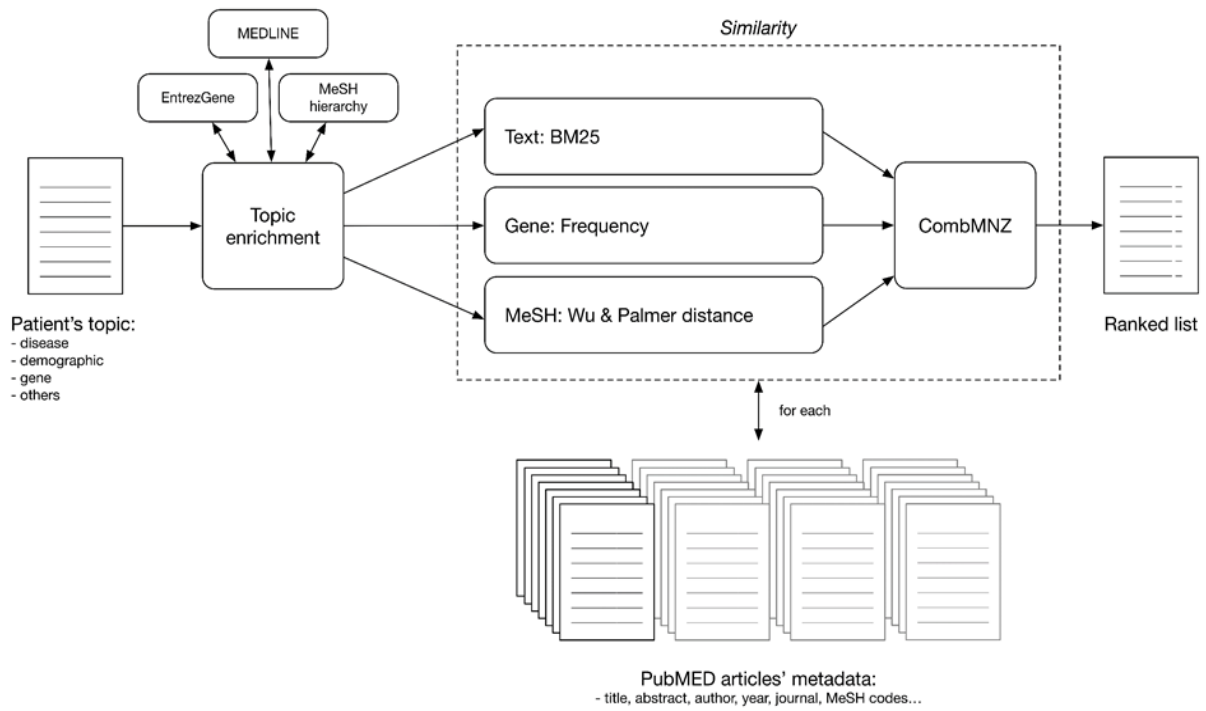

Figure 2: System architecture for ontology-based approach (biomedical article retrieval)
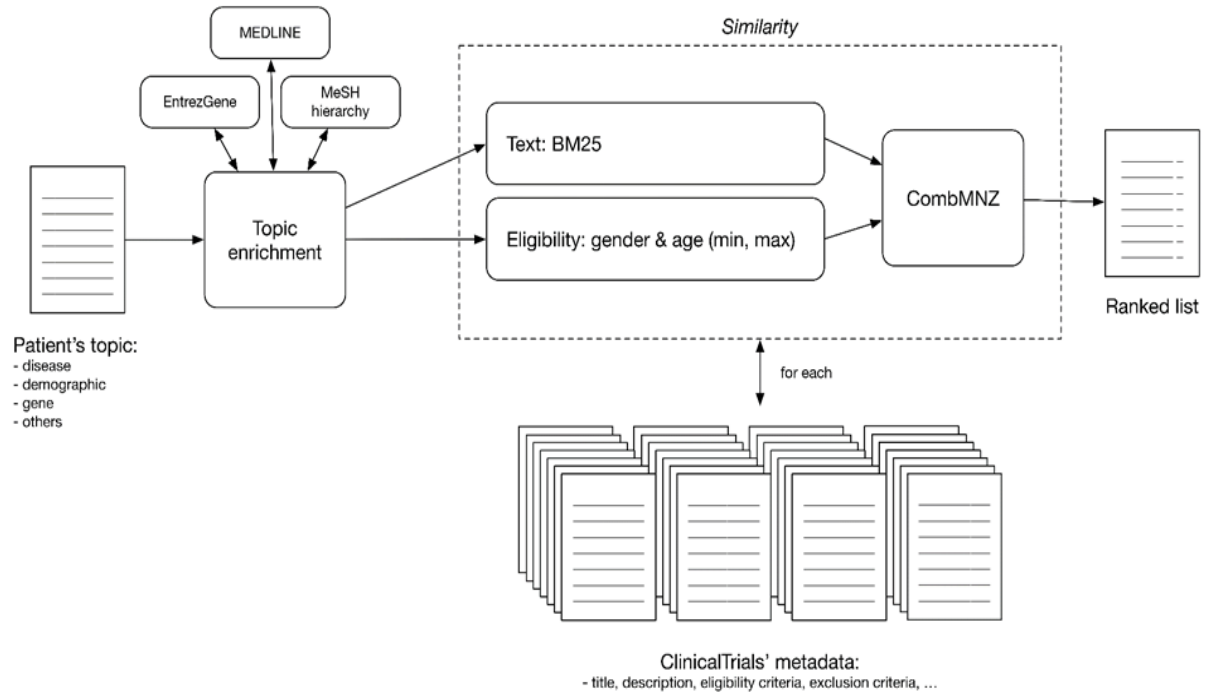

Figure 3: System architecture for ontology-based approach (clinical trial matching)

It also splits each "disease and other" concepts and links them to the nearest concept in the MeSH hierarchy.

After these preprocessing steps, the system treats the information retrieval (IR) problem as a simple measurement of similarity between a given clinical scenario and all the filtered articles. The similarity part has been computed by measuring scores at three different levels: textual, MeSH and gene. For the textual level, BM25 is used to compute similarity by taking into account the textual data available in the documents. We used the text of the title, the abstract, and the MeSH terms for the articles (Task 1). On the other hand, we used the disease names, the genes and their synonyms for the trials (Task 2).

For the MeSH level, a tree-path distance measure (Wu and Palmer, 1994) is used on the MeSH tree paths. We reward treatment-related articles by adding two extra concepts to any clinical scenario: 1) Chemicals and Drugs (MeSH: D); 2) Analytical, Diagnostic and Therapeutic techniques, and Equipment (MeSH: E). The MeSH score is used only for the articles (Task 1).

Finally, for the gene level, we compute the percentage of genes in the clinical scenario which have a correspondence with the genes in the articles. The scores are finally combined using the CombMNZ method (He and Wu, 2008). Figure 2 and Figure 3 show the basic architectures of this approach in terms of biomedical article retrieval and clinical trial matching, respectively.
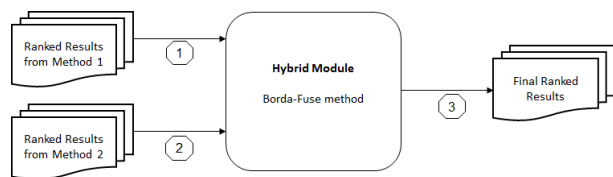

Figure 4: Hybrid system architecture

2.3 Hybrid System

We build a hybrid system based on the above two approaches using the Borda-Fuse method (Aslam and Montague, 2001) by setting a preference for the strict rule-based approach to have a higher rank in the combined ranked results. Figure 4 shows the hybrid system architecture.

3 Experimental Setup

3.1 Data

The test dataset comprises 30 clinical scenarios (called *topics*)[5] with structured patient scenarios including information related to *disease*, *genetic variants*, *demographics*, and other *relevant factors*. The dataset was curated by MD Anderson precision oncologists, and hence, included relevant information about cancer patients such that participant systems can retrieve pertinent biomedical articles and clinical trials.

3.2 Corpus

About $27M$ abstracts from a January 2017 snapshot of PubMed biomedical publications was made available by the TREC Precision Medicine track organizers this year. In addition to this, over $70K$ abstracts obtained from American Association for Cancer Research (AACR) and American Society of Clinical Oncology (ASCO) proceedings were provided for biomedical article retrieval. For the clinical trial matching task, $241K$ clinical trials from an April 2017 snapshot of ClinicalTrials.gov database[6] were provided.

3.3 Run Description

For the biomedical article retrieval task, we submitted five runs as follows: 1) *pms_run1*: considers the second approach as described in Section 2.2 without including the extra abstracts from AACR and ASCO, 2) *pms_run2_abs*: considers the first approach as described in Section 2.1 with less strict rules that allow for partial word-level matching between topical concepts and abstracts, 3) *pms_run3_abs*: uses the hybridization algorithm as described in Section 2.3 to combine the results of *pms_run2_abs* and *pms_run5_abs*, 4) *pms_run4_abs*: considers the first approach as described in Section 2.1 with stricter rules that require exact phrase-level matching between topical concepts and abstracts, and 5) *pms_run5_abs*: considers the second approach as described in Section 2.2 and extends the first run by adding the extra abstracts from AACR and ASCO.

For the clinical trial matching task, we submitted five runs as follows: 1) *pms_run1_tri*: considers the second approach as described in Section 2.2 using various internal parameters and normalization factors, 2) *pms_run2_tri*: considers the first approach as described in Section 2.1 with less strict rules that allow partial word-level matching between topical concepts and trials, 3) *pms_run3_tri*: uses the hybridization algorithm as described in Section 2.3 to combine the results of *pms_run2_tri* and *pms_run5_tri*, 4) *pms_run4_tri*: considers the first approach as described in Section 2.1 with stricter rules that allow for exact phrase-level matching between topical concepts and trials, and 5) *pms_run5_tri*: considers the second approach as described in Section 2.2 using various other internal parameters and normalization factors.

3.4 Evaluation and Analysis

The evaluation of the Precision Medicine track was conducted using the standard TREC evaluation measures for ad-hoc information retrieval tasks (Yilmaz et al., 2008; Voorhees, 2014). The highest ranked biomedical articles and clinical trials were sampled and judged by medical domain experts according to each of the four topic dimensions (disease, gene, demographic, and other), where each of them could correspond to 4 categories (e.g., a disease can be an "exact", "more general", "more specific", or "not disease" match)[7].

The main measures for biomedical article retrieval task are inferred normalized discounted cumulative gain (infNDCG), precision at R where R is the number of known relevant documents (R-prec), and precision at 10 documents (Prec (10)). Figure 5, Figure 6, and Figure 7 show the overall scores of our runs for biomedical article retrieval across all the topics as compared to the median and best scores for the submitted automatic runs. The reported results show that our biomedical article retrieval systems perform equal to or better than the *median* scores for 80.0% of the topics across all evaluation measures.

Table 1 shows the comparative results across our five submitted runs for biomedical article retrieval. The run *"pms_run5_abs"* has the best score, and in general, the runs using the ontology-based approach

---

[7] http://www.trec-cds.org/relevance_guidelines.pdf

have better scores than the runs using the strict rule matching-based approach.

Run	infNDCG	Prec (10)	R-prec
pms_run1	0.300	0.467	0.161
pms_run2_abs	0.181	0.253	0.127
pms_run3_abs	0.357	0.497	0.227
pms_run4_abs	0.266	0.383	0.159
pms_run5_abs	**0.407**	**0.530**	**0.262**

Table 1: Comparison across five runs (biomedical article retrieval)

For the clinical trial matching task, the judgment sets were created using depth-15 pools, and precision at 5 documents (Prec (5)), precision at 10 documents (Prec (10)), and precision at 15 documents (Prec (15)) are used for evaluation.

Figure 8, Figure 9, and Figure 10 show the overall scores of our runs for clinical trial matching across all topics as compared to the median and best scores for the submitted automatic runs. The reported results show that our clinical trial matching systems perform equal to or better than the *median* scores for 86.7% of the topics across all evaluation measures.

Table 2 shows the comparative results across our five submitted runs for clinical trial matching. Our run *"pms_run5_tri"* has the best score, and in general, the runs using the ontology-based approach show better performance than the runs using the strict rule matching-based approach.

Run	Prec (5)	Prec (10)	Prec (15)
pms_run1_tri	0.372	0.324	0.287
pms_run2_tri	0.186	0.145	0.108
pms_run3_tri	0.345	0.252	0.218
pms_run4_tri	0.055	0.055	0.044
pms_run5_tri	**0.455**	**0.348**	**0.317**

Table 2: Comparison across five runs (clinical trial matching)

Overall, our ontology-based approach has better results than the strict rule matching-based approach. For biomedical article retrieval task, the result of the hybrid system (*pms_run3_abs*) improved from the results for both strict rule matching-based approach (*pms_run2_abs*) and ontology-based approach (*pms_run1_abs*).

From Table 3 to Table 6, we show the scores of best runs of the top 5 teams for both biomedical article task and clinical trial matching task (Roberts et al., 2017). From these results we can see that our team (noted as **prna-mit-suny**) is consistently ranked within top 5 teams in all the evaluation metrics for the biomedical literature retrieval task (from Table 3 to Table 5) and obtains the 2nd highest precision (P@5) score for the clinical trial matching task (Table 6).

Team	Score
UTDHLTRI	0.459
BiTeM	0.409
imi_mug	0.409
UD_GU_BioTM	0.408
prna-mit-suny	**0.397**

Table 3: Top 5 teams for 29 judged topics of the biomedical article retrieval task (infNDCG scores)

Team	Score
UTDHLTRI	0.299
imi_mug	0.274
BiTeM	0.267
prna-mit-suny	**0.259**
UD_GU_BioTM	0.250

Table 4: Top 5 teams for 29 judged topics of the biomedical article retrieval task (R-prec scores)

Team	Score
UD_GU_BioTM	0.631
UTDHLTRI	0.621
imi_mug	0.617
BiTeM	0.535
prna-mit-suny	**0.521**

Table 5: Top 5 teams for 29 judged topics of the biomedical article retrieval task (P@10 scores)

4 Conclusion

In this paper, we described our participation in the TREC 2017 Precision Medicine track. The systems presented are tailored specifically for the two tasks and use strict-matching rules, textual similarity measures, and ontologies. Our best run ranked 2nd

Team	Score
UD_GU_BioTM	0.550
prna-mit-suny	**0.471**
udel	0.457
NaCTeM	0.457
NOVASearch	0.450

Table 6: Top 5 teams for 28 judged topics of the clinical trial matching task (P@5 scores)

(highest P@5) for the clinical trial matching task and consistently ranked within the top 5 teams in all the evaluation metrics for the biomedical literature retrieval task.

References

J. A. Aslam and M. Montague. 2001. Models for Metasearch. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284.

V. Datla, S. A. Hasan, A. Qadir, K. Lee, Y. Ling, J. Liu, and O. Farri. 2017. Automated Clinical Diagnosis: The Role of Content in Various Sections of a Clinical Document. In *IEEE-BIBM International Workshop on Biomedical and Health Informatics*.

S. A. Hasan, Y. Ling, J. Liu, and O. Farri. 2015. Using Neural Embeddings for Diagnostic Inferencing in Clinical Question Answering. In *Proceedings of the Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*.

Daqing He and Dan Wu. 2008. Toward a robust data fusion for document retrieval. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on*, pages 1–8. IEEE.

B. Obama. 2015. President Obama Speaks on the Precision Medicine Initiative: goo.gl/NtrUYA , January.

K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, and S. Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *TREC*.

E. M. Voorhees. 2014. The Effect of Sampling Strategy on Inferred Measures. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1119–1122.

Z. Wu and M. Palmer. 1994. Verbs Semantics and Lexical Selection. In *ACL*.

E. Yilmaz, E. Kanoulas, and J. A. Aslam. 2008. A Simple and Efficient Sampling Method for Estimating AP
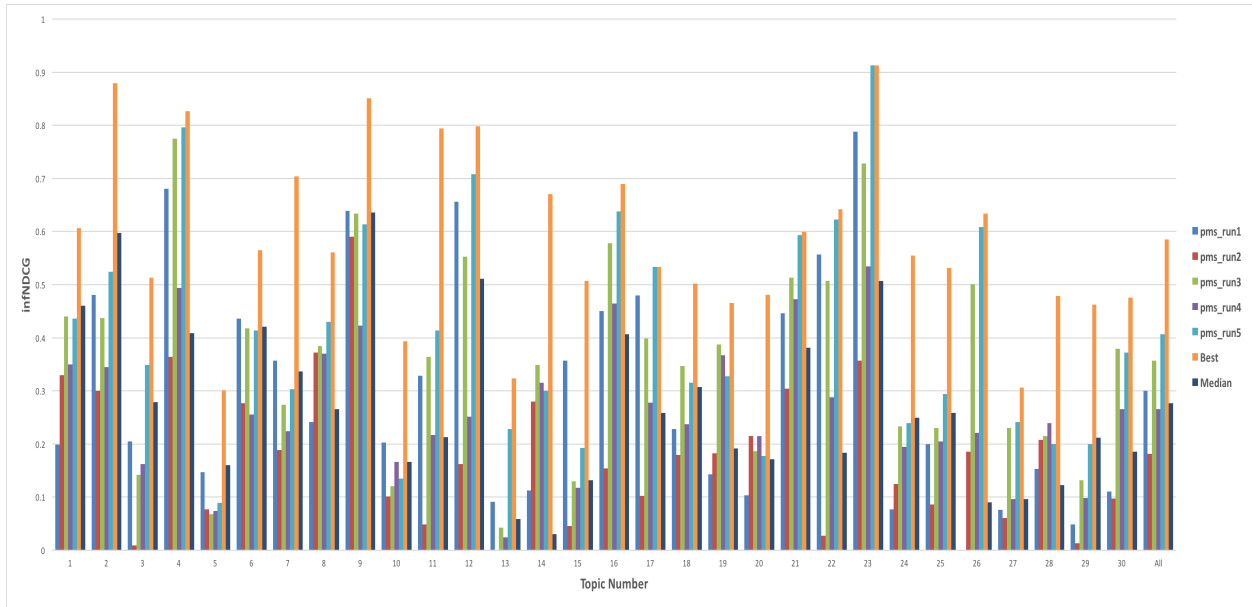
Figure 5: infNDCG scores for each topic (biomedical article retrieval)
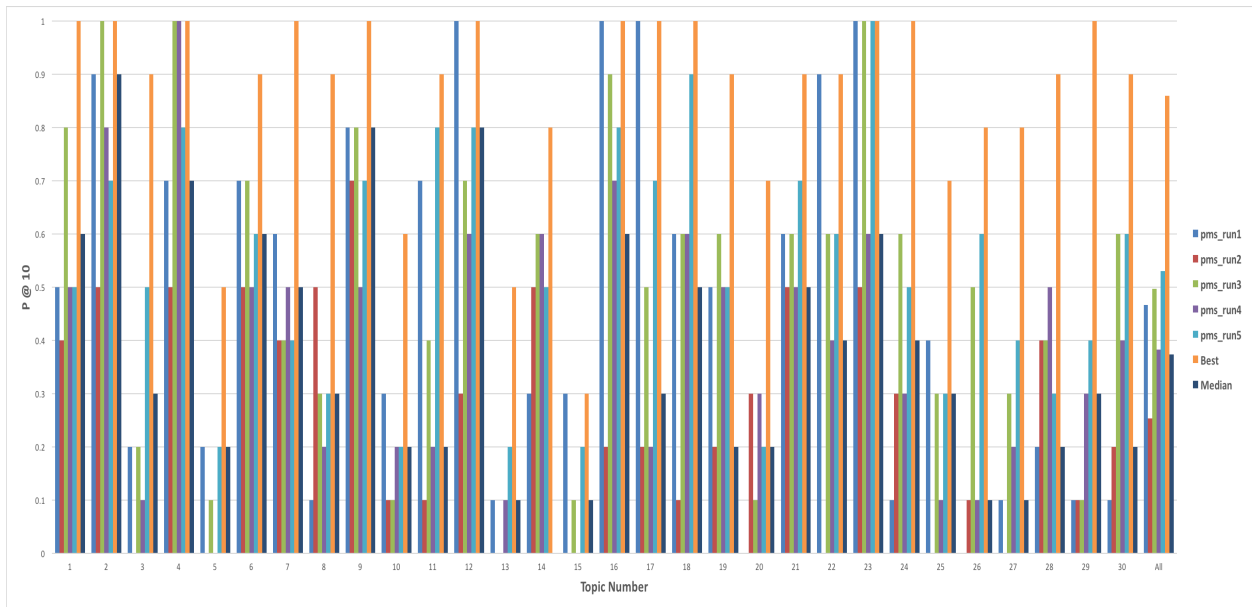

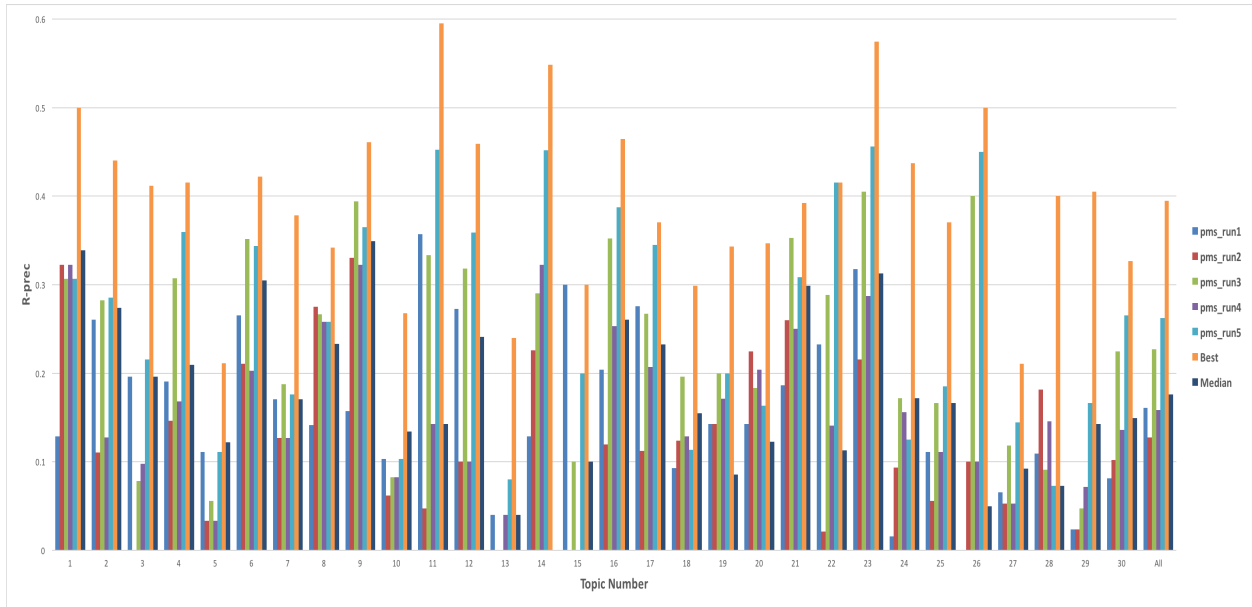Figure 6: Prec(10) scores for each topic (biomedical article retrieval)

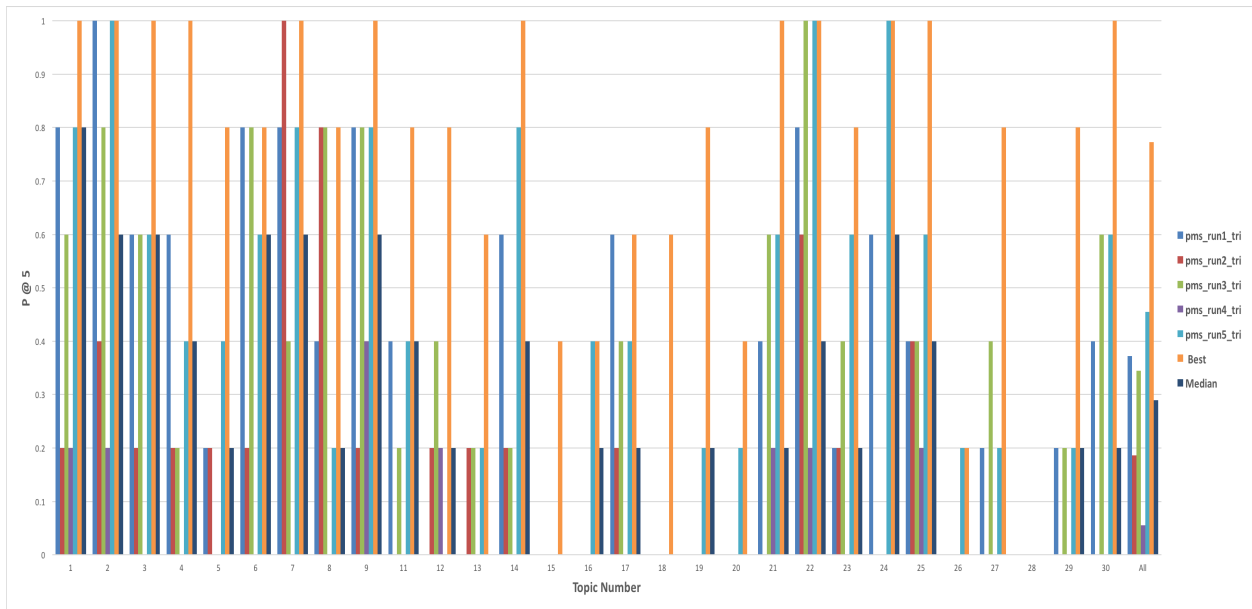Figure 7: R-prec scores for each topic (biomedical article retrieval)

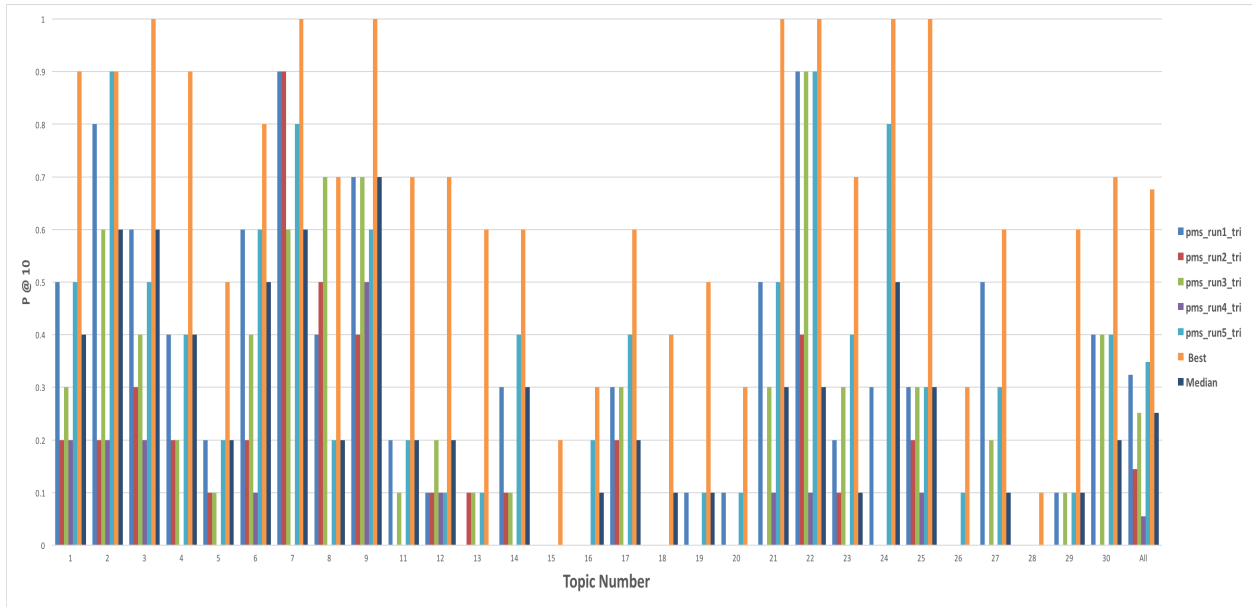Figure 8: Prec(5) scores for each topic (clinical trial matching)

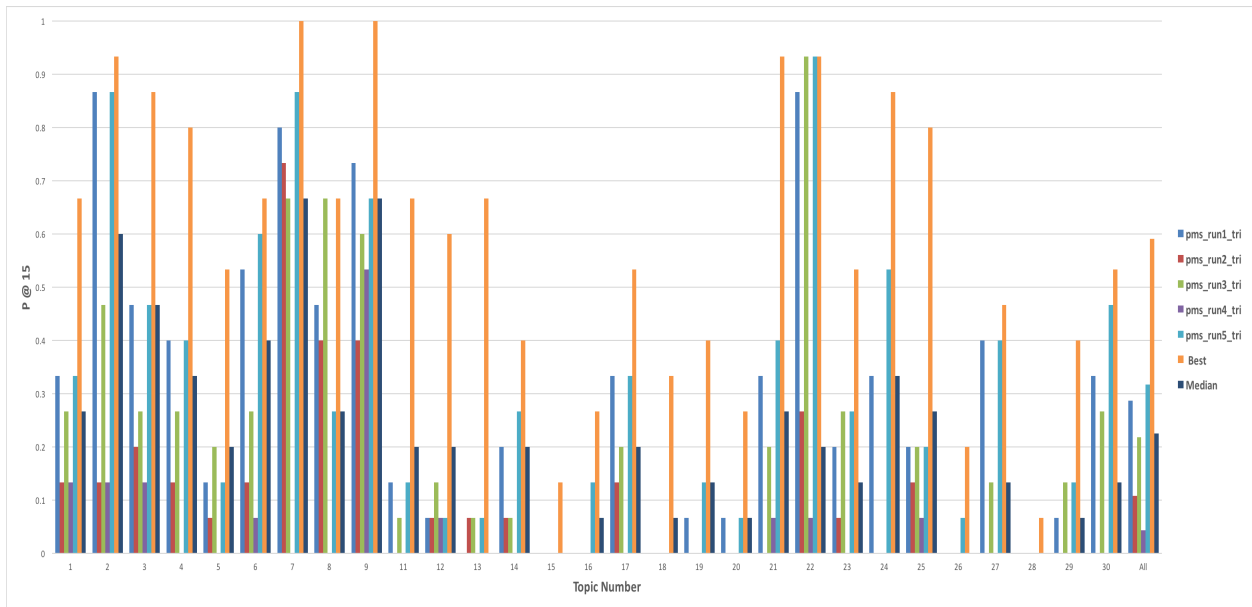Figure 9: Prec(10) scores for each topic (clinical trial matching)

Figure 10: Prec(15) scores for each topic (clinical trial matching)

and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 603–610.