Overview of the TREC 2015 Clinical Decision Support Track

Kirk Roberts, Matthew S. Simpson
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine,
National Institutes of Health, Bethesda, MD

Ellen Voorhees
National Institute of Standards and Technology, Gaithersburg, MD

William R. Hersh
Department of Medical Informatics & Clinical Epidemiology,
Oregon Health and Science University, Portland, OR

1 Introduction

In making clinical decisions, physicians often seek out information about how to best care for their patients. Information relevant to a physician can be related to a variety of clinical tasks such as (i) determining a patient's most likely diagnosis given a list of symptoms, (ii) determining if a particular test is indicated for a given situation, and (iii) deciding on the most effective treatment plan for a patient having a known condition. In some cases, physicians can find the information they seek in published biomedical literature. However, given the volume of the existing literature and the rapid pace at which new research is published, locating the most relevant and timely information for a particular clinical need can be a daunting and time-consuming task. In order to make biomedical information more accessible and to meet the requirements for the meaningful use of electronic health records (EHRs), a goal of modern clinical decision support systems is to anticipate the needs of physicians by linking EHRs with information relevant for patient care.

The goal of the 2015 TREC Clinical Decision Support (CDS) track was to evaluate biomedical literature retrieval systems for providing answers to generic clinical questions about patient cases. Short case reports, such as those published in biomedical articles and used in medical lectures, acted as idealized representations of medical records. A case report typically describes a challenging medical case. It is organized as a well-formed narrative summarizing the pertient portions of the patient's medical record. Given a case, participants were challenged with retrieving full-text biomedical articles relevant for answering questions related to one of three generic clinical information needs. The three needs were: Diagnosis (i.e., "*What is this patient's diagnosis?*"), Test ("*What diagnostic test is appropriate for this patient?*"), and Treatment ("*What treatment is appropriate for this patient?*"). Retrieved articles were judged relevant if they provided information of the specified type useful for the given case. The assessment was performed by physicians with training in biomedical informatics. The evaluation of individual submissions followed standard TREC procedures.

The 2015 CDS track differed from the 2014 CDS track (Simpson et al., 2014) by offering two tasks. Task A mirrored the 2014 CDS track, only with 30 new topics/cases. Task B used the same topics from Task A, but included the patient diagnosis for the Test and Treatment topics. Since the diagnosis was not guaranteed to be written in the case (consistent with how physicians often write cases in practice), we theorized that providing the diagnosis may improve retrieval systems by (a) providing additional relevant information if the diagnosis is not stated in the case, or (b) emphasizing a key piece of information in the case if the diagnosis is stated.

In total, 36 participating teams submitted 178 runs combined across Tasks A & B.

In the remainder of this overview paper we describe the document collection (Section 2) and topics (Section 3) provided to the participants. We then describe the evaluation (Section 4) of the retrieval results and summarize the results (Section 5) on the tasks.

2 Documents

The full-text article collection used for the track was the same as the 2014 CDS track collection. This eased the burden on returning participants and allowed all participants to evaluate their approach for both sets of topics on the same document collection.

The collection was a snapshot of the open access subset of PubMed Centeral (PMC)[1]. PMC is an online digital database of freely available full-text biomedical literature. The snapshot was obtained on January 21, 2014, containing a total of 733,138 articles. The full text of each article is represented as an NXML file (XML encoded using the U.S. National Library of Medicine (NLM) Journal Archiving and Interchange Tag Library)[2]. Images and other supplemental materials were also available. Each article in the collection is identified by a unique number (PMCID) that was used for run submissions. The PMCID of an article is specified by the `<article-id>` element within its NXML file. To make processing the document collection easier for the participants, each article file in the collection was renamed according to the article's PMCID. For example, an article with PMCID 3148967 was renamed `3148967.nxml`. The articles were available for download in 4 file bundles containing all 733,138 articles in the snapshot.

3 Topics

The topics for the track were medical case narratives created by expert topic developers at NLM that served as idealized representations of actual medical records. The case narratives described information such as a patient's medical history, the patient's current symptoms, tests performed by a physician to diagnose the patient's condition, the patient's eventual diagnosis, and finally, the steps taken by a physician to treat the patient.

In order to simulate the actual information needs of physicians, topic creators manually labeled the case narratives they constructed according to the three retrieval categories (Diagnosis, Test, Treatment). A case narrative labeled "diagnosis", for example, required participants of the track to retrieve PMC articles a physician would find useful for determining the diagnosis of the patient described in the report. Similarly, for a case narrative labeled "treatment", participants retrieved articles that would suggest to a physician the best treatment plan for the condition exhibited by the patient described in the report. Finally, participants retrieved for "test" case narratives articles that would suggest relevant interventions that a physician might undertake in diagnosing or treating the patient. When constructing the case-based topics, the topic creators were careful to omit information related to the question type. For example, a "diagnosis" report might have contained information pertaining to a patient's treatments and tests, but not the patient's diagnosis. In doing so, we hoped to more accurately mimic real clinical scenarios. The topic creators produced 10 topics for each of the 3 topic types for a total of 30 topics. All topics were vetted by an additional physician (not one of the topic creators) to ensure consistency, improve clarity, and reduce redundancy between topics.

In addition to annotating the topics according to the type of clinical information required, we also provided two versions of the case narratives. The topic "descriptions" contained a complete account of the patients' visits, including details such as their vital statistics, drug dosages, etc., whereas the topic "summaries" were simplified versions of the narratives that contained less irrelevant information. A topic's description and its summary were functionally equivalent: the set of relevant documents was identical for each version. However, we provided the summary versions of the case narratives for participants who were not interested in or equipped for processing the detailed descriptions.

The use of either the description or summary consistuted Task A of the track. Task B allowed participants to utilize an additional field: the working diagnosis. Topic creators provided this information for the Test and Treatment topics. Usually, in creating these cases, the topic creators have a specific diagnosis in

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] http://jats.nlm.nih.gov/archiving/versions.html

Topic 1 – Diagnosis
Description: A 44 yo male is brought to the emergency room after multiple bouts of vomiting that has a "coffee ground" appearance. His heart rate is 135 bpm and blood pressure is 70/40 mmHg. Physical exam findings include decreased mental status and cool extremities. He receives a rapid infusion of crystalloid solution followed by packed red blood cell transfusion and is admitted to the ICU for further care.
Summary: A 44-year-old man with coffee-ground emesis, tachycardia, hypoxia, hypotension and cool, clammy extremities.

Topic 11 – Test
Description: A 56-year old Caucasian female complains of being markedly more sensitive to the cold than most people. She also gets tired easily, has decreased appetite, and has recently tried home remedies for her constipation. Physical examination reveals hyporeflexia with delayed relaxation of knee and ankle reflexes, and very dry skin. She moves and talks slowly.
Summary: A 56-year old Caucasian female presents with sensitivity to cold, fatigue, and constipation. Physical examination reveals hyporeflexia with delayed relaxation of knee and ankle reflexes, and very dry skin.
Diagnosis: Hypothyroidism

Topic 21 – Treatment
Description: A 32-year-old male presents to your office complaining of diarrhea, abdominal cramping and flatulence. Stools are greasy and foul-smelling. He also has loss of appetite and malaise. He recently returned home from a hiking trip in the mountains where he drank water from natural sources. An iodine-stained stool smear revealed ellipsoidal cysts with smooth, well-defined walls and 2+ nuclei.
Summary: A 32-year-old male presents with diarrhea and foul-smelling stools. Stool smear reveals protozoan parasites.
Diagnosis: Giardiasis

Table 1: Topics 1, 11, and 21 from the 2015 track.

mind, even if it is not clearly stated as such in the case narrative. In Task B, participants use the diagnosis field in addition to either the description or summary. Table 1 shows examples of the case-based topics used for the track.

The topics were provided to the participants in XML format. Topic numbers were specified using the `number` attribute of each `<topic>` element and topic types (i.e., diagnosis, test, and treatment) were specified with the `type` attribute. The topic description is given in the `<description>` element and the topic summary is given in the `<summary>` element. Table 2 shows Topic 11 (from Table 1) in this format.

In order to make the results of the track more meaningful, we required that participants use only all topic descriptions or only all topic summaries for any given run submission. Participants were free to submit multiple runs so that they could experiment with the different representations. Participants were encouraged to indicate on their run submission forms which version of the topics they used.

4 Evaluation

The evaluation of the track followed standard TREC evaluation procedures for ad hoc retrieval tasks. Participants were allowed to submit in `trec_eval` format a maximum of three automatic or manual runs per topic, each consisting of a ranked list of up to one thousand PMCIDs. The assessment was performed by physicians, most of whom were biomedical informatics students in the Department of Medical Informat-

```
<topic number="11" type="test">
   <description>
      A 56-year old Caucasian female complains of being markedly more sensitive to the cold than most
      people.  She also gets tired easily, has decreased appetite, and has recently tried home
      remedies for her constipation.  Physical examination reveals hyporeflexia with delayed
      relaxation of knee and ankle reflexes, and very dry skin.  She moves and talks slowly.
   </description>
   <summary>
      A 56-year old Caucasian female presents with sensitivity to cold, fatigue, and constipation.
      Physical examination reveals hyporeflexia with delayed relaxation of knee and ankle reflexes,
      and very dry skin.
   </summary>
   <diagnosis>Hypothyroidism</diagnosis>
</topic>
```

Table 2: XML format for topic 11.

ics and Clinical Epidemiology at Oregon Health & Science University. (A few were physicians from other sites.)

The assessors were instructed to judge articles as either "definitely relevant", "not relevant", or "possibly relevant". For a document to be judged definitely relevant to a given topic, it had to provide information of the specified type (i.e., diagnosis, test, and treatment) and provide information relevant to the particular patient described in the topic. The assessors were encouraged to not view a retrieved article as providing a "correct answer" to the generic clinical question posed by the topic, but were instead instructed to judge a document relevant if there was a reasonable chance a physician might find the article useful having seen the patient described in the topic. Documents were judged not relevant if they either did not provide information of the specified type or they were not topical to the patient. Finally an article was judged possibly relevant if an assessor believed it was not immediately informative on its own, but that it may be relevant in the context of a broader literature review.

Runs were scored according to precision at 10 (P@10), R-precision (R-prec) and two inferred retrieval measures: inferred normalized discounted cumulative gain (infNDCG) and inferred average precision (infAP). See Yilmaz et al. (2008) for more details about the inferred measures. Inferred measures are used as a means of getting more accurate estimates of a run's quality than is likely possible with traditional measures when judging a relatively small number of documents.

The runs were sampled following an effective sampling strategy (Voorhees, 2014) for computing inferred measures. In particular, judgment sets were created using two strata: all documents retrieved in ranks 1-20 by any run in union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 21-100 by some run. For the evaluation reported here, the measures were computed by conflating the possibly relevant and definitely relevant sets into a single relevant set. The exception to this is the infNDCG measure, which makes use of the different relevance grades. Hence, the primary metric for comparing the retrieval submissions was infNDCG.

5 Results

A total of 36 participating teams submitted 178 accepted runs. Teams could submit up to 3 runs for Task A and 3 runs for Task B. Over half the participants submitted the maximum of 6 runs, while every participant submitted at least 2. A total of 154 fully-automatic runs were submitted, while 24 manual runs were submitted. There were 103 submissions for Task A (92 automatic), with 75 submissions for Task B (62 automatic). Table 3 lists the participating teams and their number of submissions. The total number of a participants is an increase from last year, which had 26 participants and 105 runs (though only 5 runs/participant was possible last year).

Tables 5-8 provide summarizing statistics across the automatic and manual runs for Tasks A and B. For each of the 30 topics, the tables give the best, median, and worst scores achieved by the participants. Topics 1-10 were of type Diagnosis, topics 11-20 were of type Test, and topics 21-30 were of type Treatment.

Team ID	Affiliation	# Runs			
		Task A		Task B	
		A	M	A	M
CBIA_VT	Center for Business Intelligence and Analytics, Virginia Tech	3	0	3	0
cbnu	Chonbuk National University	3	0	3	0
CL_CAMB	Computer Laboratory, University of Cambridge	3	0	3	0
DA_IICT	Dhirubhai Ambani Institute of Informantion & Communication Technology	2	0	0	0
DBNET_AUEB	Athens University of Economics and Business	2	0	2	0
DUTH	Democritus University of Thrace	2	1	3	0
ECNU	East China Normal University	3	0	3	0
EMSE	Mines Saint-Etienne	3	0	0	0
EPBRN	University of New South Wales, Australia	3	0	0	0
FDUDMIIP	School of Computer Science, Fudan University	2	1	1	2
Foreseer	University of Michigan	1	2	1	2
FORTH_ICS_ISL	Foundation for Research and Technology, Institute of Computer Science	3	0	3	0
GRIUM	GRIUM	2	0	0	0
Hipocrates15	California State University San Marcos	2	1	0	0
HITSJ	Language Technology Research Center of Harbin Institute of Technology	2	1	2	1
hltcoe	JHU Human Language Technology Center of Excellence	3	0	3	0
KISTI	Korea Institute of Science and Technology Information	3	0	3	0
LAMDA	Ajou University	3	0	3	0
LIMSI	Labo d'Informatique pour la Mecanique et les Sciences de l'Ingenieur	3	0	3	0
LIST_LUX	Luxembourg Institute of Science and Technology	3	0	1	0
NOVASEARCH	Universidade Nova de Lisboa	3	0	3	0
NU_UU_UNC	Northwestern University, University of Utah and UNC	1	2	2	1
OHSU	Oregon Health & Science University	1	2	1	2
PKUICST	Peking University	3	0	0	0
prna	Philips Research North America	3	0	3	0
SCIAITeam	Siena College Institute for Artificial Intelligence	3	0	0	2
SIBtex	SIBtex / BiTeM	3	0	0	0
SIBtex2	SIBtex / BiTeM	3	0	0	0
SNUMedinfo	Seoul National University	3	0	3	0
Sortinghat	International Institute of Information Technology Bangalore	1	0	1	0
TUW	Vienna University of Technology	3	0	3	0
udel	University of Delaware	3	0	3	0
UTDHLTRI	The University of Texas at Dallas	3	0	3	0
UWM_UO	University of Wisconsin-Milwaukee	3	0	2	1
WaterlooClarke	University of Waterloo	3	0	0	0
wsu_ir	Wayne State University	2	1	1	2
total		92	11	62	13

Table 3: Participating teams and submitted runs. (A = automatic, M = manual)

Figures 1-6 present the box-and-whiskers plots for this data across tasks and evaluation metrics.

Table 4 contains the top 5 teams in each task, split by automatic and manual systems, as well as the median and mean infNDCG scores. These teams used a variety of approaches. The top automatic approaches to Task A included: (1) Wayne State (Balaneshin-kordan et al., 2015), who utilized Markov Random Fields built from concepts in the description and top retrieved documents; (2) Luxembourg (Ben Abacha and Khelifi, 2015), who combine different information retrieval models along with semantic annotations from DBpedia; (3) Cambridge (Cummins, 2015), who utilized a Pólya urn language model; (4) ECNU (Song et al., 2015), who utilized a learning-to-rank random forest model; and (5) Delaware (Alsulmi et al., 2015), who customized term extraction based on MetaMap (Aronson and Lang, 2010) concept semantic types. The top automatic approaches to Task B included: (1) ECNU (Song et al., 2015), who combined several information retrieval models (BM25, PL2, BB2); (2) Vienna University of Technology (Palotti and Hanbury, 2016), who utilize selective MetaMap extraction similar to the Delaware team; (3) Seoul National University, who did not submit a notebook paper; (4) LIMSI (D'hondt et al., 2015), who utilized MeSH in a variety of ways; and (5) Cambridge (Cummins, 2015), who used the same approach as Task A but included the diagnosis field along with the summary. The top manual approaches for Tasks A and B generally limited the manual interventions to two areas: (a) Wayne State (Balaneshin-kordan et al., 2015), Northwestern (Stöber et al., 2015), and Fudan (You et al., 2015) all manually filtered keywords, while (b) Northwestern and Fudan also manually added a diagnosis.

Overall, compared to the 2014 results, the infNDCG scores have risen. The average median infNDCG

Participant	infNDCG
Task A - Automatic	
Wayne State Univ. [wsu_ir]	0.2939
Luxembourg IST [LIST_LUX]	0.2894
Univ. of Cambridge [CL_CAMB]	0.2823
East China Normal Univ. [ECNU]	0.2680
Univ. of Delaware [udel]	0.2676
Median	0.2288
Mean	0.2099
Task A - Manual	
Wayne State Univ. [wsu_ir]	0.3109
Northwestern/Utah/UNC [NU_UU_UNC]	0.3019
Univ. of Michigan [Foreseer]	0.2954
Fudan Univ. [FDUDMIIP]	0.2689
Median	0.2504
Mean	0.2496
Demo. Univ. of Thrace [DUTH]	0.2318
Task B - Automatic	
East China Normal Univ. [ECNU]	0.3821
Vienna Univ. of Tech. [TUW]	0.3616
Seoul National Univ. [SNUMedinfo]	0.3611
LIMSI [LIMSI]	0.3507
Univ. of Cambridge [CL_CAMB]	0.3471
Median	0.3095
Mean	0.2870
Task B - Manual	
Fudan Univ. [FDUDMIIP]	0.3809
Wayne State Univ. [wsu_ir]	0.3690
Univ. of Michigan [Foreseer]	0.3535
Northwestern/Utah/UNC [NU_UU_UNC]	0.3255
Median	0.3212
Harbin Inst. of Tech. [HITSJ]	0.3168
Mean	0.2842

Table 4: Top 5 results by infNDCG. Mean and Median scores are based on the best submitted run for each participant.

score for the 2014 task was 0.15141, while the average median infNDCG for Task A was 0.20384 (Task B is not directly comparable to 2014). Whether this is a result of easier topics or improved systems cannot be assessed at this time.

When assessing the differences between the two tasks, Task B had much higher scores. The average median infNDCG on Task B was 0.27937 (compared to 0.20384 for Task A). This could likely mean that providing the diagnosis does improve overall system performance. However, there are potential confounders here as well. Fewer participants submitted runs for Task B, and it might be possible that the average Task A-only participant had inferior approachs to the average of those participants who submitted to both tasks. It is difficult to judge within the individual participants who submitted to both tasks as well, since they might have used entirely different approaches to Task A and Task B. Despite this, it seems quite encouraging that providing the diagnosis improves retrieval results.

6 Conclusion

2015 was the second year of the Clinical Decision Support track. The goal of the track is to inform the creation of clinical decision support systems that bring scientific evidence (in the form of biomedical literature) to the point-of-care. Participants were provided with simulated case narratives, and challenged with finding relevant scientific articles to address questions of diagnosis, testing, and treatment. Participation in the track was extraordinary, increasing from 26 participating teams in 2014 to 36 teams this year. In addition to having the previous year's topics for system improvement, the track provided the diagnosis (Task B) to test whether this improved retrieval. While not conclusive, there are signs that retrieval systems improved from the 2014 track to this year. It also appears that providing the diagnosis for Test and Treatment topics

provides a boost to performance. It is hoped that in future evaluations, system performance will continue to increase.

Acknowledgements

We would like to thank Swapna Abhyankar, Dina Demner-Fushman, Bryan Hendrickson, Fabricio Kury, Laritza Rodriguez, and Raymonde Uy for their assistance in creating topics. This work was partially supported by U.S. National Library of Medicine of the National Institutes of Health under both award number 1K99LM012104 and the intramural research program.

References

Alsulmi, M., Sabhnani, K., and Carterette, B. (2015). University of Delaware at TREC 2015. In *Proceedings of the 2015 Text Retrieval Conference*.

Aronson, A. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17:229–236.

Balaneshin-kordan, S., Kotov, A., and Xisto, R. (2015). WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources. In *Proceedings of the 2015 Text Retrieval Conference*.

Ben Abacha, A. and Khelifi, S. (2015). LIST at TREC 2015 Clinical Decision Support Track: Question Analysis and Unsupervised Result Fusion. In *Proceedings of the 2015 Text Retrieval Conference*.

Cummins, R. (2015). Clinical Decision Support with the SPUD Language Model. In *Proceedings of the 2015 Text Retrieval Conference*.

D'hondt, E., Grau, B., and Zweigenbaum, P. (2015). LIMSI @ 2015 Clinical Decision Support Track. In *Proceedings of the 2015 Text Retrieval Conference*.

Palotti, J. and Hanbury, A. (2016). TUW @ TREC Clinical Decision Support Track 2015. In *Proceedings of the 2015 Text Retrieval Conference*.

Simpson, M. S., Voorhees, E., and Hersh, W. (2014). Overview of the TREC 2014 Clinical Decision Support Track. In *Proceedings of the 2014 Text Retrieval Conference*.

Song, Y., He, Y., Hu, Q., and He, L. (2015). ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval. In *Proceedings of the 2015 Text Retrieval Conference*.

Stöber, J., Heale, B. S., Kim, H., Fulghum, K., Raja, K., Fiol, G. D., and Jonnalagadda, S. R. (2015). Concept based Information Retrieval for Clinical Case Summaries. In *Proceedings of the 2015 Text Retrieval Conference*.

Voorhees, E. M. (2014). The effect of sampling strategy on inferred measures. In *Proceedings of the 37th Annual ACM International Conference on Research and Development in Information Retrieval*, pages 1119–1122.

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual ACM International Conference on Research and Development in Information Retrieval*, pages 603–610.

You, R., Zhou, Y., Peng, S., and Zhu, S. (2015). FDUMedSearch at TREC 2015 Clinical Decision Support Track. In *Proceedings of the 2015 Text Retrieval Conference*.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0352	0.0079	0.0000	0.2763	0.1209	0.0000	0.1836	0.0870	0.0000	0.7000	0.2000	0.0000
2	0.0488	0.0098	0.0000	0.1521	0.0508	0.0000	0.1667	0.0667	0.0000	0.3000	0.1000	0.0000
3	0.1473	0.0193	0.0000	0.9646	0.3119	0.0000	0.4134	0.1760	0.0028	1.0000	0.5000	0.0000
4	0.0765	0.0381	0.0002	0.6321	0.4014	0.0211	0.3195	0.2236	0.0160	0.9000	0.7000	0.0000
5	0.0512	0.0054	0.0000	0.1945	0.0498	0.0000	0.1852	0.0370	0.0000	0.2000	0.1000	0.0000
6	0.0872	0.0051	0.0000	0.3185	0.0447	0.0000	0.2400	0.0533	0.0000	0.5000	0.1000	0.0000
7	0.0495	0.0124	0.0000	0.2961	0.1455	0.0000	0.2667	0.1333	0.0000	0.7000	0.2000	0.0000
8	0.1658	0.0882	0.0000	0.5016	0.3780	0.0000	0.3906	0.3047	0.0000	1.0000	0.7000	0.0000
9	0.1167	0.0269	0.0000	0.2944	0.1202	0.0000	0.2979	0.1064	0.0000	0.7000	0.3000	0.0000
10	0.1217	0.0307	0.0000	0.3803	0.1603	0.0000	0.3457	0.1852	0.0000	0.5000	0.2000	0.0000
11	0.1305	0.0025	0.0000	0.5713	0.0682	0.0000	0.3739	0.0315	0.0000	1.0000	0.1000	0.0000
12	0.0895	0.0112	0.0000	0.2247	0.0777	0.0000	0.2258	0.0645	0.0000	0.3000	0.1000	0.0000
13	0.1094	0.0457	0.0000	0.4271	0.2466	0.0000	0.3198	0.2267	0.0000	1.0000	0.6000	0.0000
14	0.1560	0.0556	0.0000	0.4476	0.2253	0.0000	0.3256	0.1395	0.0000	0.6000	0.2000	0.0000
15	0.0519	0.0104	0.0006	0.4200	0.1641	0.0253	0.2650	0.1066	0.0055	1.0000	0.4000	0.0000
16	0.0779	0.0465	0.0000	0.6085	0.4676	0.0000	0.4788	0.3328	0.0044	1.0000	0.9000	0.0000
17	0.1149	0.0564	0.0029	0.7029	0.4384	0.0684	0.4366	0.3717	0.0236	1.0000	0.7000	0.2000
18	0.0047	0.0001	0.0000	0.0849	0.0080	0.0000	0.0667	0.0074	0.0000	0.1000	0.0000	0.0000
19	0.0441	0.0066	0.0000	0.2404	0.0692	0.0000	0.1591	0.0682	0.0000	0.5000	0.1000	0.0000
20	0.0661	0.0005	0.0000	0.3146	0.0160	0.0000	0.1169	0.0130	0.0000	0.5000	0.0000	0.0000
21	0.1617	0.0342	0.0000	0.5138	0.1965	0.0000	0.4097	0.2014	0.0000	0.8000	0.3000	0.0000
22	0.0751	0.0473	0.0014	0.8322	0.5690	0.0678	0.4757	0.3171	0.0157	1.0000	0.9000	0.1000
23	0.1296	0.0523	0.0000	0.3902	0.2459	0.0000	0.3670	0.2385	0.0000	0.8000	0.3000	0.0000
24	0.0328	0.0020	0.0000	0.1729	0.0294	0.0000	0.1500	0.0333	0.0000	0.3000	0.1000	0.0000
25	0.0422	0.0000	0.0000	0.0679	0.0000	0.0000	0.0625	0.0000	0.0000	0.1000	0.0000	0.0000
26	0.3056	0.1908	0.0000	0.6652	0.5085	0.0000	0.6364	0.4416	0.0000	1.0000	0.8000	0.0000
27	0.1253	0.0036	0.0000	0.4185	0.0476	0.0000	0.2750	0.0250	0.0000	0.3000	0.0000	0.0000
28	0.1138	0.0035	0.0000	0.3462	0.0316	0.0000	0.3396	0.0377	0.0000	0.7000	0.0000	0.0000
29	0.6875	0.2240	0.0000	0.8796	0.4072	0.0000	0.7297	0.3514	0.0000	1.0000	0.8000	0.0000
30	0.3560	0.2045	0.0000	0.8574	0.5149	0.0000	0.6434	0.4651	0.0000	1.0000	0.9000	0.0000

Table 5: Per-topic summary results for 92 automatic runs on Task A.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0478	0.0244	0.0031	0.2773	0.2011	0.0583	0.2029	0.1208	0.0725	0.7000	0.5000	0.2000
2	0.0816	0.0252	0.0000	0.2647	0.0841	0.0000	0.2333	0.1000	0.0000	0.4000	0.2000	0.0000
3	0.1134	0.0739	0.0207	0.9405	0.6571	0.2697	0.3939	0.3128	0.1453	1.0000	0.8000	0.5000
4	0.0650	0.0363	0.0104	0.5356	0.4214	0.2078	0.2939	0.1885	0.1150	0.9000	0.6000	0.2000
5	0.0580	0.0064	0.0000	0.3429	0.0485	0.0000	0.1111	0.0000	0.0000	0.2000	0.0000	0.0000
6	0.1457	0.0154	0.0000	0.3888	0.1173	0.0000	0.3067	0.1333	0.0000	0.6000	0.2000	0.0000
7	0.0384	0.0200	0.0074	0.2995	0.1887	0.0805	0.2267	0.1267	0.0533	0.5000	0.2000	0.1000
8	0.1197	0.0709	0.0044	0.4625	0.3005	0.0673	0.3984	0.2500	0.0781	0.8000	0.6000	0.0000
9	0.0871	0.0563	0.0000	0.2507	0.1680	0.0000	0.2340	0.1489	0.0000	0.7000	0.4000	0.0000
10	0.0763	0.0258	0.0036	0.2601	0.1523	0.0210	0.2840	0.1852	0.0247	0.4000	0.1000	0.0000
11	0.2300	0.0230	0.0000	0.7039	0.2484	0.0000	0.5090	0.1757	0.0000	1.0000	0.4000	0.0000
12	0.0926	0.0321	0.0004	0.2686	0.1657	0.0212	0.2258	0.0968	0.0000	0.3000	0.1000	0.0000
13	0.1152	0.0641	0.0098	0.3974	0.3121	0.0913	0.3023	0.1919	0.0349	0.9000	0.7000	0.3000
14	0.1282	0.0282	0.0070	0.3805	0.1299	0.0620	0.2791	0.1628	0.0233	0.5000	0.2000	0.0000
15	0.0662	0.0128	0.0030	0.5318	0.1983	0.0952	0.2377	0.1230	0.0410	0.9000	0.5000	0.1000
16	0.0725	0.0515	0.0092	0.5960	0.5089	0.2777	0.4190	0.2920	0.0964	1.0000	0.9000	0.6000
17	0.0683	0.0403	0.0231	0.5049	0.3808	0.2444	0.4189	0.3127	0.0973	0.8000	0.6000	0.2000
18	0.1812	0.0013	0.0000	0.6568	0.0768	0.0000	0.3630	0.0370	0.0000	0.9000	0.1000	0.0000
19	0.0234	0.0042	0.0000	0.1614	0.0584	0.0000	0.1364	0.0568	0.0000	0.4000	0.1000	0.0000
20	0.0650	0.0080	0.0008	0.2440	0.0721	0.0192	0.1948	0.0649	0.0260	0.5000	0.0000	0.0000
21	0.0775	0.0329	0.0136	0.2914	0.1854	0.0886	0.2639	0.2014	0.0694	0.8000	0.3000	0.1000
22	0.0781	0.0565	0.0302	0.8109	0.7204	0.4795	0.4223	0.3014	0.0895	1.0000	1.0000	0.5000
23	0.1237	0.0674	0.0029	0.3813	0.2443	0.0313	0.3578	0.1927	0.0367	0.8000	0.4000	0.1000
24	0.2075	0.0031	0.0004	0.4164	0.0349	0.0156	0.3833	0.0333	0.0167	0.5000	0.0000	0.0000
25	0.0222	0.0000	0.0000	0.1740	0.0000	0.0000	0.0625	0.0000	0.0000	0.1000	0.0000	0.0000
26	0.3059	0.1653	0.0128	0.6571	0.4551	0.1294	0.6104	0.4026	0.1299	1.0000	0.8000	0.1000
27	0.1440	0.0127	0.0003	0.3802	0.0838	0.0084	0.3500	0.1250	0.0000	0.4000	0.1000	0.0000
28	0.0509	0.0018	0.0000	0.1899	0.0294	0.0000	0.1698	0.0189	0.0000	0.5000	0.0000	0.0000
29	0.6617	0.3856	0.0132	0.8545	0.5442	0.0778	0.7432	0.5000	0.0811	1.0000	0.8000	0.2000
30	0.3458	0.1535	0.0393	0.6785	0.4193	0.2364	0.6512	0.3488	0.2248	1.0000	0.9000	0.3000

Table 6: Per-topic summary results for 11 manual runs on Task A.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0825	0.0074	0.0000	0.4382	0.1122	0.0000	0.2029	0.0870	0.0000	1.0000	0.2000	0.0000
2	0.0454	0.0053	0.0000	0.2068	0.0407	0.0000	0.1667	0.0333	0.0000	0.3000	0.0000	0.0000
3	0.1230	0.0225	0.0000	0.9932	0.3319	0.0000	0.3939	0.1927	0.0028	1.0000	0.5000	0.0000
4	0.0765	0.0401	0.0000	0.6730	0.4174	0.0000	0.3291	0.2236	0.0000	0.9000	0.7000	0.0000
5	0.1061	0.0058	0.0000	0.3541	0.0542	0.0000	0.2963	0.0370	0.0000	0.3000	0.1000	0.0000
6	0.0546	0.0053	0.0000	0.2549	0.0489	0.0000	0.1867	0.0533	0.0000	0.6000	0.1000	0.0000
7	0.0469	0.0166	0.0000	0.2781	0.1575	0.0000	0.2667	0.1400	0.0000	0.6000	0.2000	0.0000
8	0.1658	0.0906	0.0004	0.5016	0.3955	0.0064	0.3906	0.3125	0.0078	1.0000	0.7000	0.0000
9	0.1162	0.0425	0.0000	0.3051	0.1429	0.0000	0.2979	0.1489	0.0000	0.7000	0.3000	0.0000
10	0.0988	0.0307	0.0000	0.3290	0.1703	0.0000	0.3457	0.1605	0.0000	0.4000	0.2000	0.0000
11	0.2352	0.0562	0.0000	0.7518	0.3424	0.0000	0.5946	0.2613	0.0000	1.0000	0.8000	0.0000
12	0.1377	0.0414	0.0000	0.3387	0.1691	0.0000	0.2258	0.0968	0.0000	0.4000	0.1000	0.0000
13	0.1128	0.0631	0.0000	0.4427	0.3074	0.0000	0.3547	0.2384	0.0000	1.0000	0.7000	0.0000
14	0.1626	0.0836	0.0000	0.4734	0.3036	0.0000	0.3488	0.2093	0.0000	0.6000	0.3000	0.0000
15	0.0695	0.0392	0.0000	0.5393	0.3770	0.0000	0.3224	0.2268	0.0082	1.0000	0.8000	0.0000
16	0.0731	0.0451	0.0011	0.7608	0.5691	0.0652	0.4934	0.2788	0.0117	1.0000	1.0000	0.0000
17	0.1159	0.0588	0.0000	0.6955	0.4393	0.0000	0.4218	0.3569	0.0059	1.0000	0.7000	0.0000
18	0.1643	0.0021	0.0000	0.6609	0.0418	0.0000	0.3926	0.0444	0.0000	0.8000	0.1000	0.0000
19	0.0936	0.0132	0.0000	0.3827	0.1119	0.0000	0.2500	0.1023	0.0000	0.6000	0.2000	0.0000
20	0.2127	0.0586	0.0000	0.4566	0.2587	0.0000	0.4675	0.2597	0.0000	0.9000	0.4000	0.0000
21	0.3914	0.1104	0.0000	0.8053	0.4425	0.0000	0.5069	0.3403	0.0000	1.0000	0.6000	0.0000
22	0.0745	0.0634	0.0001	0.9003	0.7711	0.0194	0.5071	0.3595	0.0016	1.0000	1.0000	0.0000
23	0.1266	0.0788	0.0000	0.4068	0.2944	0.0000	0.4037	0.2936	0.0000	0.7000	0.4000	0.0000
24	0.1974	0.0133	0.0000	0.4490	0.0752	0.0000	0.3833	0.0667	0.0000	0.7000	0.2000	0.0000
25	0.2534	0.0000	0.0000	0.4228	0.0000	0.0000	0.4375	0.0000	0.0000	0.5000	0.0000	0.0000
26	0.3014	0.2407	0.0000	0.6646	0.5532	0.0000	0.6494	0.5130	0.0000	1.0000	0.9000	0.0000
27	0.1616	0.0181	0.0000	0.4198	0.1315	0.0000	0.3500	0.1000	0.0000	0.7000	0.1000	0.0000
28	0.1639	0.0546	0.0000	0.3991	0.2076	0.0000	0.4528	0.2264	0.0000	0.8000	0.4000	0.0000
29	0.6950	0.3829	0.0000	0.8828	0.5893	0.0000	0.7162	0.5405	0.0000	1.0000	0.9000	0.0000
30	0.3526	0.2086	0.0000	0.8574	0.5244	0.0000	0.6357	0.4651	0.0000	1.0000	0.9000	0.0000

Table 7: Per-topic summary results for 62 automatic runs on Task B.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0589	0.0097	0.0022	0.3016	0.1511	0.0494	0.2029	0.0870	0.0048	0.7000	0.3000	0.1000
2	0.2474	0.0194	0.0000	0.4464	0.0598	0.0000	0.4000	0.0667	0.0000	0.4000	0.2000	0.0000
3	0.1388	0.0444	0.0062	0.8307	0.5302	0.1415	0.3966	0.2263	0.0196	1.0000	0.9000	0.4000
4	0.0717	0.0456	0.0109	0.5431	0.4363	0.1325	0.3131	0.2556	0.0256	0.9000	0.8000	0.3000
5	0.0580	0.0078	0.0003	0.3429	0.0512	0.0136	0.0741	0.0370	0.0000	0.1000	0.1000	0.0000
6	0.0918	0.0108	0.0000	0.3402	0.0769	0.0000	0.2667	0.0800	0.0000	0.5000	0.2000	0.0000
7	0.0446	0.0232	0.0004	0.3097	0.2006	0.0159	0.2400	0.1267	0.0067	0.7000	0.2000	0.1000
8	0.1146	0.0767	0.0044	0.4415	0.3630	0.0673	0.3750	0.2891	0.0547	0.8000	0.6000	0.0000
9	0.0976	0.0730	0.0002	0.2655	0.1748	0.0082	0.2553	0.1702	0.0000	0.7000	0.6000	0.0000
10	0.0664	0.0292	0.0031	0.2412	0.1431	0.0212	0.3210	0.1728	0.0123	0.4000	0.2000	0.0000
11	0.2368	0.1404	0.0063	0.7203	0.6194	0.1104	0.5225	0.4414	0.0180	1.0000	0.9000	0.4000
12	0.0829	0.0465	0.0000	0.2865	0.2038	0.0000	0.2258	0.1290	0.0000	0.3000	0.2000	0.0000
13	0.1360	0.0655	0.0110	0.4464	0.3251	0.0913	0.3605	0.2209	0.0349	1.0000	0.7000	0.3000
14	0.1495	0.0670	0.0070	0.4148	0.2665	0.0620	0.3023	0.2093	0.0698	0.6000	0.3000	0.0000
15	0.0662	0.0483	0.0018	0.5077	0.3797	0.0420	0.3115	0.2268	0.0109	0.9000	0.8000	0.2000
16	0.0768	0.0520	0.0066	0.6917	0.5345	0.1320	0.4934	0.2088	0.0146	1.0000	1.0000	0.5000
17	0.0843	0.0582	0.0084	0.5805	0.4919	0.1038	0.4130	0.3510	0.0236	0.9000	0.8000	0.4000
18	0.1438	0.0047	0.0000	0.5505	0.0995	0.0000	0.3556	0.0815	0.0000	0.7000	0.1000	0.0000
19	0.0363	0.0093	0.0000	0.2216	0.0953	0.0000	0.1591	0.0909	0.0000	0.3000	0.1000	0.0000
20	0.0840	0.0482	0.0000	0.3106	0.2206	0.0000	0.2987	0.2078	0.0000	0.4000	0.2000	0.0000
21	0.3255	0.0757	0.0154	0.7336	0.2834	0.0873	0.4722	0.3056	0.0417	0.9000	0.5000	0.3000
22	0.0693	0.0568	0.0074	0.8143	0.7350	0.1519	0.4835	0.3579	0.0157	1.0000	1.0000	0.8000
23	0.1143	0.0706	0.0029	0.3647	0.2621	0.0313	0.3761	0.2752	0.0367	0.7000	0.4000	0.1000
24	0.1989	0.0159	0.0000	0.3932	0.0804	0.0000	0.3333	0.0667	0.0000	0.4000	0.1000	0.0000
25	0.1353	0.0006	0.0000	0.3054	0.0217	0.0000	0.3750	0.0000	0.0000	0.4000	0.0000	0.0000
26	0.3042	0.2271	0.0128	0.6658	0.5467	0.1294	0.6494	0.3766	0.0649	1.0000	0.9000	0.1000
27	0.1632	0.0101	0.0019	0.4185	0.0885	0.0159	0.3750	0.0500	0.0000	0.6000	0.1000	0.0000
28	0.3090	0.0461	0.0000	0.4930	0.2242	0.0000	0.3962	0.1698	0.0000	0.6000	0.3000	0.0000
29	0.6359	0.4775	0.0161	0.7727	0.6405	0.0836	0.7162	0.5000	0.0811	1.0000	0.8000	0.2000
30	0.3493	0.1384	0.0393	0.6900	0.3928	0.1614	0.6357	0.3256	0.0698	1.0000	0.9000	0.3000

Table 8: Per-topic summary results for 13 manual runs on Task B.

Figure 1: Task A average infNDCG results.

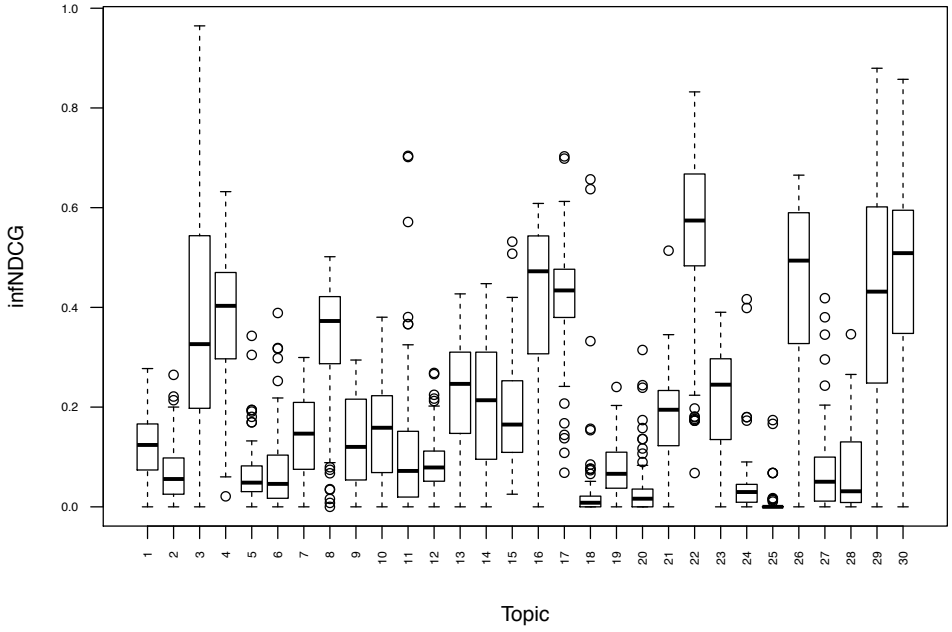

Figure 2: Task B average infNDCG results.

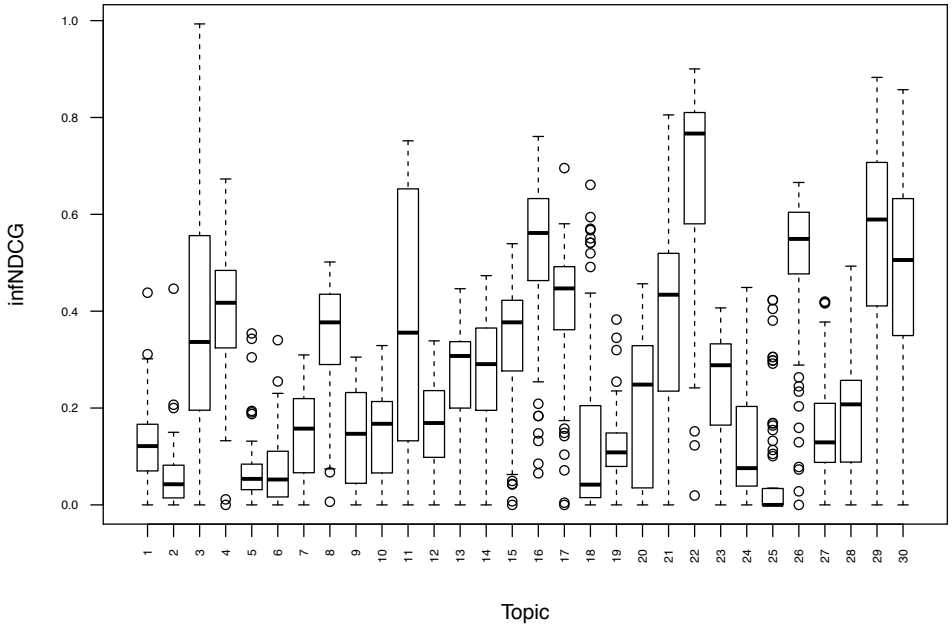

Figure 3: Task A average infAP results.

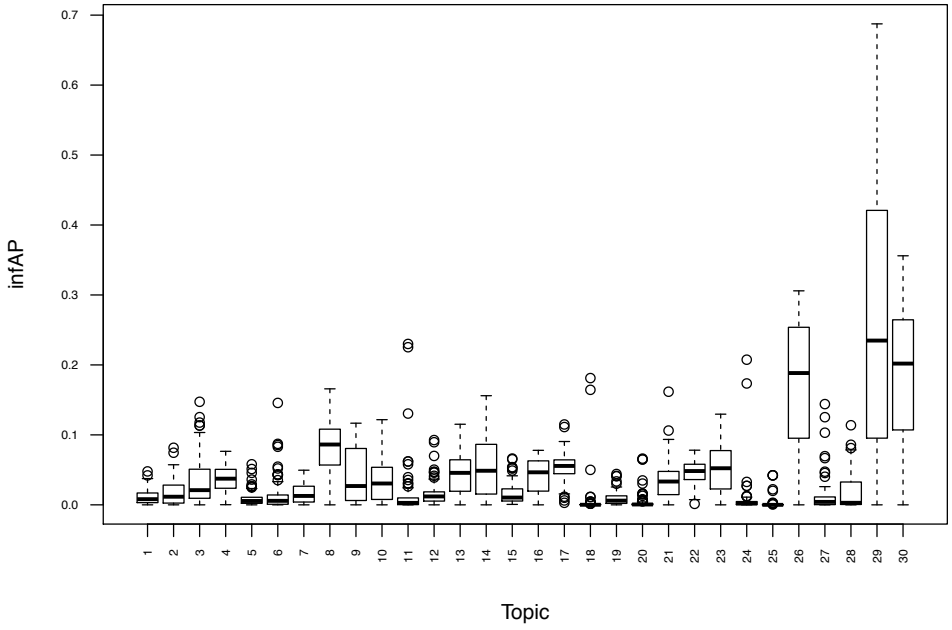

Figure 4: Task B average infAP results.

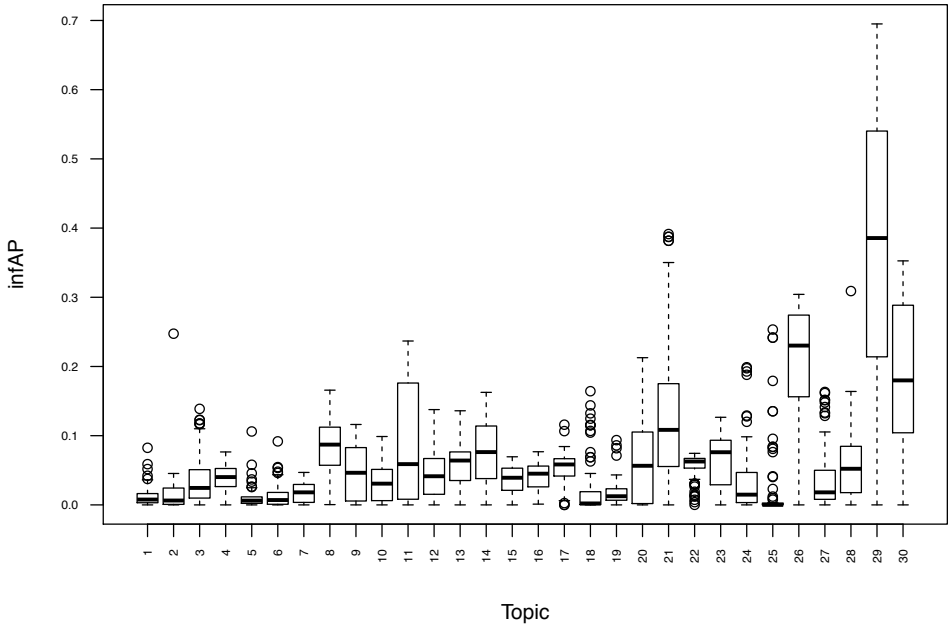

Figure 5: Task A average Precision @ 10 results.

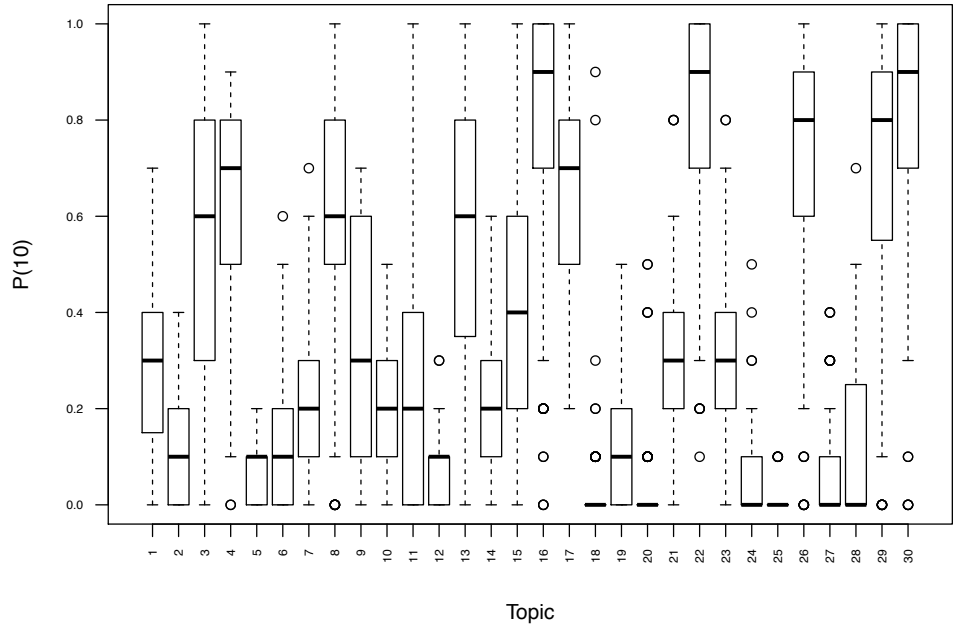

Figure 6: Task B average Precision @ 10 results.