Overview of the TREC 2016 Clinical Decision Support Track

Kirk Roberts
School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD

Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD

Ellen M. Voorhees
Information Technology Laboratory, 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 handling challenging cases, clinicians often seek out information to make better decisions in patient care. Typically, these information sources combine clinical experience with scientific medical research in a process known as evidence-based medicine (EBM). 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. Finding the most relevant and recent research, however, can be quite challenging due to the volume of scientific literature and the pace at which new research is published. As such, the time-consuming nature of information seeking means that most clinician questions go unanswered (Ely et al., 2005).

In order to better enable access to the scientific literature in the clinical setting, research is necessary to evaluate information retrieval methods that connect clinical notes with the published literature. The TREC Clinical Decision Support (CDS) track simulates the information retrieval requirements of such systems to encourage the creation of tools and resources necessary for their implementation. In 2014 and 2015, the CDS tracks used simulated patient cases presented as if they were typical case reports used in medical education. However, in an actual electronic health record (EHR), patient notes are written in a much different manner, notably with terse language and heavy use of abbreviations and clinical jargon. To address the challenge specific to EHR notes, the 2016 CDS track used de-identified notes for actual patients. This enabled participants to experiment with a realistic topic/query and develop methods to handle the challenging nature of clinical text. For a given EHR note, participants were challenged with retrieving full-text biomedical articles relevant for answering questions related to one of three generic clinical information needs: 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 using a 3-point scale: relevant, partially relevant, not relevant.

In total, 26 participating teams submitted 115 runs.

In the remainder of this overview paper we describe the background and track history (Section 2), document collection (Section 3), and topics (Section 4) provided to the participants. We then describe the evaluation (Section 5) of the retrieval results and summarize the results (Section 6) on the tasks.

2 Track Background & History

The CDS Track has been heavily inspired by the TREC Genomics (Hersh and Voorhees, 2009) and Medical Records (Voorhees and Hersh, 2012) tracks in addition to the medical case-based retrieval track of ImageCLEF (Seco de Herrera et al., 2013), all of which are no longer active. All of these tracks have demonstrated significant interest in the problem of medical ad hoc retrieval.

Without a reusable, de-identified collection of medical records, the TREC 2014 and 2015 Clinical Decision Support tracks (Simpson et al., 2014; Roberts et al., 2015) proposed the use of short case reports, such as those published in biomedical articles, as idealized representations of actual medical records. For a given case report, participants of the tracks were challenged with retrieving full-text biomedical articles relevant for answering questions related to several types of clinical information needs. The 2014 and 2015 tracks were very popular: in 2014, 26 teams participated, submitting 105 total runs, while in 2015, 36 teams participated, submitting 178 total runs.

The current iteration, the 2016 CDS track, strived for a balance of continuity and progress toward a more realistic task. First, the document collection was updated to a more recent snapshot of PubMed Central (from 730k to 1.25 million full-text articles). Second, and most notably, the topics were made more realistic for the clinical setting. While the case reports used in previous iterations contain much the same information as the original EHR clinical note, the note is filled with abbreviations and jargon and written in a telegraphic style. It is unreasonable to expect a busy clinician to re-write her note in a manner similar to a case report simply to get better search results. Instead, information retrieval systems should be tasked with operating on the note directly, as this minimizes the time burden for a clinician. As such, the 2016 track provided topics in the form of an actual, de-identified EHR clinical note. To ensure continuity, the case report description abstracted from the note was also provided, but participants were encouraged to submit runs using the note itself as the topic query.

3 Documents

The full-text article collection used for the track was drawn from the open access subset of PubMed Central (PMC)[1], as in 2014 and 2015. However, an updated snapshot was used since the original snapshot was from January 2014. PMC is an online digital database of freely available full-text biomedical literature. 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 1.25 million articles in the snapshot.

4 Topics

The topics for the track were nursing admission notes obtained from the MIMIC-III database (Johnson et al., 2016). The MIMIC team gave the track special permission to make 30 notes publicly available without the need for a Data Use Agreement (DUA). MIMIC-III notes have already been automatically de-identified, but to ensure maximal privacy protection, manual de-identification was performed as well. Further, only notes from deceased patients were used.

MIMIC records were obtained from Boston-area Intensive Care Units (ICU). Since the patients are undergoing critical care, their medical problems are necessarily less diverse than the topics from previous years. However, there is still a wide diversity of ICU patients. To help ensure a diversity of cases selected as topics, the notes were automatically clustered using K-means with 30 clusters according to the patient's

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[2] http://jats.nlm.nih.gov/archiving/versions.html

Topic 1 – Diagnosis
Note:
78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**]
(transferred to nursing home for rehab on [**12-8**] after several falls out
of bed.) He was then readmitted to [**Hospital6 1749**] on
[**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?.
There was a question whether he had a small MI; he reportedly had a
small NQWMI. He improved with diuresis and was not intubated.
.
Yesterday, he was noted to have a melanotic stool earlier this evening
and then approximately 9 loose BM w/ some melena and some frank blood
just prior to transfer, unclear quantity.
Description:
78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.
Summary:
A 78 year old male presents with frequent stools and melena.

Table 1: Topic 1.

associated ICD-9 codes (billing codes that roughly represent the patient's problems). Topic creators were given a cluster with a selection of up to 10 admission notes. They chose one of the notes in the cluster to be the topic, discarding the rest. Where necessary, topic creators removed parts of the note to hide details of the actual diagnosis, test, or treatment performed. Topic creators also trimmed the notes to reduce overall length when necessary, removing sections that largely were not manually entered (but added automatically from structured fields within the EHR). Specifically, topic creators kept the History of Present Illness (HPI) section, which corresponds to the "case report" used in previous tracks. Once a clinical note was selected, manually de-identified, and edited where appropriate, the topic creators then constructed a case description and summary similar to the 2014 and 2015 tasks. Example topics (clinical notes, description, and case summary) are shown in Tables 1, 2, and 3.

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's note is given in the `<note>` element, the description is given in the `<description>` element, and the summary is given in the `<summary>` element. Table 4 shows Topic 1 (from Table 1) in this format.

To encourage participants to utilize the note version of the topic, at most 3 total runs (of the maximum 5 allowed) could utilize either the description or summary. That is, teams submitting the maximum of 5 allotted runs were required to use the note version on at least 2 runs. For a given run, only one version of the topic could be used (e.g., on a single run one could not utilize the note for some of the topics, the description for others, and the summary for the rest). Participants were required to indicate which version of the topic was used.

Similar to previous years, the track divided topics among three different types:

1. **Diagnosis**. This type can be interpreted as asking the question: "What is the patient's diagnosis?" This corresponds to the Ely et al. (2000) classification `Diagnosis/Cause`. A topic of this type would require participants to return PMC articles a physician would find useful for determining the diagnosis of a patient described in a case report.
2. **Treatment**. This type can be interpreted as asking the question: "What is the best treatment for this patient's condition?" and corresponds to the Ely et al. (2000) classifications `Treatment/Drugs/Indications` and `Treatment/General/Indications`. A topic of this type would require participants to return PMC articles a physician would find useful for creating the best treatment plan for the condition exhibited by the patient described in the case report.
3. **Test**. This type can be interpreted as asking the question: "What is the best intervention for diagnosing this patient's condition?" and corresponds to the Ely et al. (2000) classification `Diagnosis/Test`. A topic of this type would require participants to return PMC articles a physician would find useful for pursuing the best diagnostic procedure for the patient described in the case report.

Topic 11 – Test
Note:
Mr. [**Name13 (STitle) 5827**] is an 80yo M with dementia, CAD s/p CABG in [**3420**] (LIMA-LAD,
SVG to OM2, SVG to RPDA), then s/p CABG redo in [**3426**], then s/p 2 caths
this year with patent LIMA, totally occluded SVG to RPDA, SVG to OM2,
s/p BMS to LCX on [**1-26**] who presented to [**Hospital3 53**] Hospital
with increasing chest pain and nausea over the past few days.
.
Per report, patient has presented several times since last cathed for
recurrent angina. Admitted to [**Hospital3 **] on [**3436-4-2**] with recurrent chest pain. Ruled out for MI. Last episode of chest pressure was the morning of transfer, associated with dry heaves and belching relieved with
morphine. Pt was continued on ASA, Plavix, Statin, BBker, Imdur and
placed on Heparin gtt. Cath last [**Month (only) **] here at [**Hospital1 5**] showed a patent BMS in LCX and no new lesions. According to the
family he usually has angina once every day or two, but for the past 2
weeks he has been having angina with any minimal exertion (eg putting
on his shirt), and waking him several times per night.
Description:
A 80yo male with dementia and past history of CABG, two caths this year patent LIMA, totally occluded SVG to RPDA, SVG to OM2, s/p BMS to LCX, presents with increasing chest pain and nausea over the past few days. The patient has history of repeated episodes of recurrent chest pain with relief with morphine. Pt is on ASA, Statins, Imdur, and Heparin. Last months cath showed patent BMS in LCX and no new lesions. According to the family, the patient has increasing episodes of chest pain with minimal exertion in the last two weeks.
Summary:
80 yo male with demantia and past medical history of CABG with repeated episodes of chest pain. Admitted for severe chest pain episode.

Table 2: Topic 11.

For each type, participants can tailor their retrieval methods to retrieve information particular to that type to meet the corresponding clinical need.

5 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 five automatic or manual runs per topic, each consisting of a ranked list of up to 1,000 PMCIDs per topic. The assessment was performed by physicians, most of whom were either biomedical informatics students (in the Department of Medical Informatics and Clinical Epidemiology at Oregon Health & Science University) or postdoctoral fellows (at the Lister Hill National Center for Biomedical Communications at the U.S. National Library of Medicine).

Assessors judged articles as either "Definitely Relevant", "Possibly Relevant", or "Not 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 for computing inferred measures. See Voorhees (2014) for more information on sampling. In particular, judgment sets were created using two strata: all documents retrieved in ranks 1-15 by any run in union with a 20% sample of documents not

Topic 21 – Treatment
Note:
Mr. [**Known patient lastname 4075**] is a 63 yo man with h/o biphenotypic ALL, now Day + 32
from allogeneic SCT, who presents to clinc with one week of worsening
SOB and two days of a clear productive cough. The patient states his
SOB occured when lying flat, but not with activity. Also admitted to
chest pressure which would come and go in his left chest no related to
the SOB. Sleeps with 3 pillows (no change from baseline), denies PND;
admits to a slight increase in lower extremity edema. Admits to low
grade fevers to the 99's and crampy abdominal pain. Denies chills,
night sweats, vomiting, or diarrhea.
Assessment and Plan
Assesment: This is a 63 year-old male with a history of h/o
biphenotypic ALL, now Day + 32 from allogeneic SCT, who presents with
hypoxia, one week of worsening SOB, and two days of productive cough.
Plan:
# Hypoxia: The patient developed acute onset of hypoxia accompanied by
fever and a one day cough with sputum production. Given that the
patient is about 1 month s/p allogenic SCT the differential is broad
and would include bacterial pneumonia, viral pneumonia (CMV, flu), and
opportunistic infections including fungal infections. Patient also has
a history of CMV infection, aspergillus and Leggionare's disease and is on
posaconazole. His CXR showed an opacification of the left basilar lobe
and also right upper lobe concerning for pneumonia as well as a small
loculated right pleural effusion. Also in the differential is
noninfectious causes such as PE, CHF, or MI. US were negative for clot
and his first set of CE were negative.
Description:
A 63 yo man with h/o biphenotypic ALL, now Day + 32 from allogeneic SCT, who presents with one week of worsening SOB and two days of a clear productive cough. The patient states his SOB occured when lying flat, but not with activity. Also admitted to chest pressure which would come and go in his left chest no related to the SOB. Sleeps with 3 pillows (no change from baseline), denies PND; admits to a slight increase in lower extremity edema. Admits to low grade fevers to the 99's and crampy abdominal pain. Denies chills, night sweats, vomiting, or diarrhea. Patient also has a history of CMV infection, aspergillus and Leggionare's disease and is on posaconazole. His CXR showed an opacification of the left basilar lobe and also right upper lobe concerning for pneumonia as well as a small loculated right pleural effusion.
Summary:
A 63 year-old male with biphenotypic ALL, Day +32 after BMT, h/o CMV infection, aspergillus and Leggionare's disease, presents with acute onset of hypoxia accompanied by fever and two days of productive cough. His CXR showed an opacification of the left basilar lobe and also right upper lobe concerning for pneumonia.

Table 3: Topic 21.

retrieved in the first set that were retrieved in ranks 16-100 by some run. For the evaluation reported here, most measures were computed by combining the "Possibly Relevant" and "Definitely Relevant" sets into a single relevant set. The exception is the infNDCG measure, which makes use of the different relevance grades. For this reason, the primary metric for comparing the retrieval submissions was infNDCG.

6 Results

A total of 26 participating teams submitted 115 accepted runs. Teams could submit up to 5 runs. Over half the participants submitted the maximum of 5 runs, while every participant submitted at least 2. A total of 107 fully-automatic runs were submitted, while 8 manual runs were submitted. A total of 48 submitted runs used the original notes, 20 used the description, and 47 used the summary. Table 5 lists the participating teams and their number of submissions.

Table 6 provides the top automatic runs for each of the four measures. Tables 7-10 provide summarizing statistics across the automatic and manual runs, giving the best, median, and worst scores achieved by the participants on each topic. Topics 1-10 were of type Diagnosis, topics 11-20 were of type Test, and topics 21-30 were of type Treatment. Figures 1-8 present the box-and-whiskers plots for the top runs across the three note types and two primary evaluation metrics (infNDCG and P@10).

In comparing results for the topic representations, it appears the summary was the highest performing

```
<topic number="1" type="diagnosis">
  <note>
    78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**]
    (transferred to nursing home for rehab on [**12-8**] after several falls out
    of bed.)  He was then readmitted to [**Hospital6 1749**] on
    [**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?.
    There was a question whether he had a small MI; he reportedly had a
    small NQWMI. He improved with diuresis and was not intubated.
    .
    Yesterday, he was noted to have a melanotic stool earlier this evening
    and then approximately 9 loose BM w/ some melena and some frank blood
    just prior to transfer, unclear quantity.
  </note>
  <description>
    78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI.
Yesterday, he was noted to have a melanotic stool and then today he had approximately 9 loose BM w/
some melena and some frank blood just prior to transfer, unclear quantity.
  </description>
  <summary>
    A 78 year old male presents with frequent stools and melena.
  </summary>
</topic>
```

Table 4: XML format for topic 1.

representation when averaging the median run performance in both infNDCG (note: 0.1228, description: 0.1043, summary: 0.1859) and P@10 (note: 0.1833, description: 0.1533, summary: 0.2633). As seen in Table 6, the top runs all used the summary. Interestingly, note runs tended to outperform description runs, though this might be due to the smaller sample size (48 automatic note runs vs. 20 description runs).

Regarding topic types, the Test topics (11-20) appear to have the best results, followed by the Treatment topics (21-30), with Diagnosis (1-10) performing the worst. This pattern holds across all topic representations (note, description, summary) for these topic types. For instance, with the summary topic representations, the average median infNDCG was 0.1784 for Diagnosis, 0.1970 for Test, and 0.1823 for Treatment.

Average results appear to be down from the 2015 task (Roberts et al., 2015), but higher than the original 2014 track (Simpson et al., 2014). The drop in performance from last year could easily be due to the way in which topics were generated in terms of (a) a sample of actual notes vs. synthetic, and (b) ICU notes vs. general practice.

7 Conclusion

The 2016 track was the third iteration 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. Unlike past years, participants were provided with actual clinical cases, and challenged with finding relevant scientific articles to address questions of diagnosis, testing, and treatment. Participation in the track was excellent, totaling 26 participating teams. Results indicate that using the original clinical note itself is indeed challenging for information retrieval systems. It is hoped that future research on this dataset will improve how systems handle the variety of problems presented by clinical text.

Acknowledgements

We would like to thank Swapna Abhyankar, Fabricio Kury, Laritza Rodriguez, Raymonde Uy, and Ferdinand Dhombres for their assistance in creating topics. We would also like to thank Kate Fultz Hollis for her help with the assessment process. This work was partially supported by U.S. National Library of Medicine of the National Institutes of Health under award number R00LM012104, as well as the intramural research

		# Runs			
Team ID	Affiliation	Note	Desc	Summ	Total
cbnu	Chonbuk National University	1	-	2	3
CCNU2016TREC	Central China Normal University	3	1	1	5
CSIROmed	Commonwealth Sci. and Ind. Research Org.	2	1	2 (1)	5
DA_IICT	Dhirubhai Ambani IICT	3	1	1	5
DUTH	Democritus University of Thrace	-	-	3	3
ECNU	East China Normal University	2 (1)	-	3	5
ETH	ETH Zurich	2	1	2	5
FDUDMIIP	Fudan University	1	1	2	4
hany-miner	University of Michigan	2	2	1	5
HAUT	Henan University of Technology	1	1	1	3
IAII_PUT	Poznan University of Technology	2	2	1	5
iris	University of Pittsburgh	2 (2)	2	1	5
IRIT	Institut de Recherche en Informatique de Toulouse	-	2	-	2
MayoNLPTeam	Mayo Clinic	2 (1)	2 (1)	1	5
MERCKKGAA	MERCK KGAA	1	-	3	4
nch_risi	Nationwide Children's Hospital	2 (1)	-	3	5
NKU	Nankai University	3	1	1	5
NLM_NIH	U.S. National Library of Medicine	2	-	3	5
prna	Philips Research North America	3	1	1	5
SCIAICLTeam	Siena College	2	-	3	5
udel_fang	University of Delaware (Fang)	3	1	1	5
udel	University of Delaware (Carterett)	2	-	3	5
UIowaS	University of Iowa	1	1	1	3
UNTIIA	University of North Texas	3 (1)	-	2	5
UWM	University of Wisconsin-Milwaukee	-	-	3	3
WHUIRGroup	Wuhan University	3	-	2	5
total		48 (6)	20 (1)	47 (1)	115

Table 5: Participating teams and submitted runs. Numbers in parentheses indicate manual runs.

program of the National Library of Medicine. Furthermore, we are extremely greatful to the National Insitute of Standards and Technology (NIST) for sponsoring the track and assessments.

References

Ely, J., Osheroff, J., Chambliss, M., Ebell, M., and Rosenbaum, M. (2005). Answering Physicians' Clinical Questions: Obstacles and Potential Solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224. PMC551553.

Ely, J. W., Osheroff, J. A., Gorman, P. M., Ebell, M. H., Chambliss, M. L., Pifer, E. A., and Stavri, P. Z. (2000). A taxonomy of generic clinical questions: classification study. *BMJ*, 321:429–432. PMC27459.

Hersh, W. and Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval*, 12:1–15.

Johnson, A. E., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., , and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Roberts, K., Simpson, M. S., Voorhees, E., and Hersh, W. (2015). Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of the 2015 Text Retrieval Conference*.

Seco de Herrera, A. G., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., and Müller, H. (2013). Overview of the ImageCLEF 2013 medical tasks. In *CLEF 2013 Working Notes*.

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*.

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.

infAP				infNDCG			
Team	Run	Score	Topic	Team	Run	Score	Topic
FDUDMIIP	AutoSummary1	0.0454	S	FDUDMIIP	AutoSummary1	0.2815	S
nch_risi	SumES	0.0321	S	MERCKKGAA	MrkUmlsXgb	0.2493	S
CCNU2016TREC	CCNUSUMR1	0.0316	S	cbnu	cbnus1	0.2382	S
cbnu	cbnus1	0.0316	S	udel_fang	UDelInfoCDS5	0.2362	S
MERCKKGAA	MrkUmlsXgb	0.0315	S	ECNU	ECNUrun5	0.2334	S
ECNU	ECNUrun5	0.0313	S	DUTH	DUTHsaRPF	0.2265	S
udel_fang	UDelInfoCDS5	0.0311	S	nch_risi	SumES	0.2222	S
udel	udelSRef	0.0302	S	ETH	ETHSummRR	0.2179	S
DUTH	DUTHsaRPF	0.0302	S	CCNU2016TREC	CCNUSUMR1	0.2179	S
NKU	nkuRun1	0.0289	S	MayoNLPTeam	mayoas	0.2146	S

R-prec				P@10			
Team	Run	Score	Topic	Team	Run	Score	Topic
MERCKKGAA	MrkUmlsXgb	0.1744	S	FDUDMIIP	AutoSummary1	0.4033	S
MayoNLPTeam	mayoas	0.1659	S	MERCKKGAA	MrkUmlsXgb	0.3500	S
FDUDMIIP	AutoSummary1	0.1633	S	cbnu	cbnus1	0.3400	S
ECNU	ECNUrun1	0.1598	S	nch_risi	SumES	0.3378	S
CCNU2016TREC	AutoSummary	0.1578	S	udel_fang	UDelInfoCDS5	0.3367	S
IAII_PUT	SDPHBo1NE	0.1565	S	ECNU	ECNUrun5	0.3367	S
udel_fang	UDelInfoCDS5	0.1527	S	MayoNLPTeam	mayoas	0.3067	S
NLM_NIH	NLMrun2	0.1465	S	ETH	ETHSummRR	0.3067	S
udel	udelSRef	0.1464	S	udel	udelSRef	0.3033	S
UWM	UWM1	0.1463	S	CCNU2016TREC	AutoSummary	0.2967	S

Table 6: Top overall automatic systems (best run per participant). N = note, D = description, S = summary (*all top runs used the summary*).

Voorhees, E. M. and Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. In *Proceedings of the 11th Text REtrieval Conference*.

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.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0773	0.0060	0.0001	0.3436	0.0863	0.0049	0.2812	0.0547	0.0078	0.5000	0.2000	0.0000
2	0.1330	0.0032	0.0000	0.4999	0.0444	0.0000	0.1471	0.0294	0.0000	0.3000	0.0000	0.0000
3	0.0323	0.0053	0.0000	0.2380	0.1030	0.0000	0.1316	0.0724	0.0000	0.4000	0.2000	0.0000
4	0.0217	0.0005	0.0000	0.0588	0.0109	0.0000	0.0556	0.0000	0.0000	0.1000	0.0000	0.0000
5	0.0369	0.0008	0.0000	0.2146	0.0163	0.0000	0.1531	0.0306	0.0000	0.5000	0.0000	0.0000
6	0.0560	0.0340	0.0025	0.3064	0.2152	0.0310	0.2057	0.1702	0.0284	0.5000	0.4000	0.2000
7	0.0978	0.0481	0.0011	0.4009	0.2029	0.0250	0.2338	0.0909	0.0260	0.6000	0.3000	0.0000
8	0.0368	0.0300	0.0210	0.7369	0.6071	0.3986	0.2541	0.2026	0.0340	1.0000	0.9000	0.8000
9	0.0278	0.0193	0.0035	0.2015	0.1793	0.0388	0.1789	0.1463	0.0569	0.5000	0.2000	0.1000
10	0.1249	0.0052	0.0000	0.3527	0.0605	0.0000	0.1579	0.0000	0.0000	0.3000	0.0000	0.0000
11	0.0507	0.0263	0.0054	0.4579	0.3809	0.1105	0.2616	0.2044	0.1144	0.9000	0.6000	0.3000
12	0.0529	0.0152	0.0009	0.2757	0.1762	0.0254	0.1171	0.0901	0.0180	0.6000	0.2000	0.0000
13	0.2086	0.0031	0.0000	0.7693	0.0699	0.0000	0.1842	0.0658	0.0000	1.0000	0.2000	0.0000
14	0.0250	0.0147	0.0011	0.2107	0.1566	0.0337	0.1500	0.0667	0.0417	0.4000	0.1000	0.0000
15	0.1185	0.0069	0.0029	0.3636	0.0909	0.0500	0.1818	0.0779	0.0390	0.5000	0.2000	0.0000
16	0.1817	0.0033	0.0002	0.7892	0.0725	0.0093	0.1685	0.0543	0.0163	1.0000	0.2000	0.0000
17	0.0280	0.0039	0.0014	0.2299	0.0724	0.0341	0.2260	0.1073	0.0395	0.6000	0.1000	0.0000
18	0.0258	0.0038	0.0000	0.1784	0.0480	0.0000	0.1471	0.0588	0.0000	0.2000	0.1000	0.0000
19	0.0755	0.0158	0.0021	0.4252	0.1845	0.0528	0.1306	0.0946	0.0631	0.7000	0.2000	0.0000
20	0.0561	0.0421	0.0232	0.8930	0.6264	0.4636	0.3541	0.2128	0.0729	1.0000	1.0000	0.5000
21	0.1186	0.0054	0.0000	0.3392	0.0407	0.0000	0.2329	0.0411	0.0000	0.8000	0.1000	0.0000
22	0.0017	0.0000	0.0000	0.0256	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	0.0208	0.0038	0.0001	0.1334	0.0692	0.0043	0.1009	0.0734	0.0092	0.3000	0.1000	0.0000
24	0.0318	0.0226	0.0166	0.5859	0.4223	0.3835	0.2564	0.1804	0.0374	0.9000	0.7000	0.4000
25	0.0917	0.0209	0.0001	0.5567	0.2339	0.0051	0.3790	0.1324	0.0046	0.8000	0.4000	0.0000
26	0.1024	0.0281	0.0063	0.3896	0.1663	0.0716	0.3540	0.1681	0.0619	0.7000	0.4000	0.1000
27	0.0898	0.0256	0.0000	0.2754	0.0934	0.0000	0.1538	0.0769	0.0000	0.2000	0.1000	0.0000
28	0.0408	0.0008	0.0001	0.3659	0.0746	0.0077	0.1604	0.0425	0.0189	0.9000	0.1000	0.0000
29	0.1855	0.0443	0.0029	0.5512	0.2389	0.0462	0.3729	0.2458	0.0254	0.8000	0.2000	0.1000
30	0.0858	0.0071	0.0002	0.2472	0.0362	0.0070	0.2000	0.0500	0.0000	0.4000	0.1000	0.0000

Table 7: Per-topic results for 8 manual runs.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0723	0.0004	0.0000	0.3539	0.0118	0.0000	0.2734	0.0156	0.0000	0.6000	0.0000	0.0000
2	0.0274	0.0000	0.0000	0.1288	0.0000	0.0000	0.1176	0.0000	0.0000	0.3000	0.0000	0.0000
3	0.0236	0.0022	0.0000	0.1944	0.0582	0.0000	0.1250	0.0526	0.0000	0.3000	0.0000	0.0000
4	0.0226	0.0023	0.0000	0.0693	0.0209	0.0000	0.1111	0.0556	0.0000	0.2000	0.0000	0.0000
5	0.0112	0.0004	0.0000	0.0814	0.0101	0.0000	0.0816	0.0102	0.0000	0.3000	0.0000	0.0000
6	0.0712	0.0221	0.0003	0.3537	0.1769	0.0079	0.2695	0.1489	0.0071	0.8000	0.3000	0.0000
7	0.0448	0.0051	0.0000	0.2144	0.0540	0.0000	0.2208	0.0649	0.0000	0.5000	0.1000	0.0000
8	0.0429	0.0214	0.0022	0.7548	0.4606	0.1169	0.3279	0.1487	0.0070	1.0000	0.8000	0.0000
9	0.0540	0.0106	0.0000	0.2836	0.0933	0.0000	0.2195	0.0976	0.0000	0.4000	0.1000	0.0000
10	0.1672	0.0031	0.0000	0.3668	0.0352	0.0000	0.2105	0.0000	0.0000	0.3000	0.0000	0.0000
11	0.0505	0.0255	0.0007	0.5181	0.3294	0.0306	0.2752	0.1989	0.0082	0.9000	0.5000	0.1000
12	0.0272	0.0023	0.0000	0.2178	0.0543	0.0000	0.1622	0.0541	0.0000	0.4000	0.1000	0.0000
13	0.0063	0.0008	0.0000	0.0871	0.0283	0.0000	0.0592	0.0197	0.0000	0.3000	0.0000	0.0000
14	0.0065	0.0012	0.0000	0.1215	0.0415	0.0000	0.1000	0.0333	0.0000	0.2000	0.0000	0.0000
15	0.0354	0.0053	0.0000	0.2361	0.0818	0.0000	0.1818	0.0649	0.0000	0.4000	0.1000	0.0000
16	0.0168	0.0026	0.0000	0.2271	0.0583	0.0036	0.1141	0.0326	0.0054	0.3000	0.1000	0.0000
17	0.0881	0.0044	0.0000	0.5126	0.0753	0.0000	0.2542	0.0678	0.0000	0.8000	0.2000	0.0000
18	0.0251	0.0046	0.0000	0.1682	0.0600	0.0000	0.1324	0.0735	0.0000	0.4000	0.1000	0.0000
19	0.0346	0.0029	0.0000	0.2408	0.0642	0.0000	0.1306	0.0541	0.0000	0.7000	0.1000	0.0000
20	0.0530	0.0326	0.0017	0.8555	0.6032	0.0782	0.3723	0.2128	0.0091	1.0000	0.8000	0.2000
21	0.1543	0.0130	0.0000	0.4094	0.0810	0.0000	0.2877	0.0959	0.0000	0.4000	0.2000	0.0000
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	0.0474	0.0008	0.0000	0.2368	0.0188	0.0000	0.1193	0.0275	0.0000	0.5000	0.1000	0.0000
24	0.0332	0.0189	0.0002	0.6559	0.4119	0.0215	0.2436	0.1418	0.0039	1.0000	0.6000	0.1000
25	0.1194	0.0407	0.0000	0.7406	0.3898	0.0000	0.3653	0.2511	0.0046	0.9000	0.5000	0.0000
26	0.1675	0.0236	0.0000	0.5348	0.1571	0.0000	0.3363	0.1681	0.0000	0.9000	0.3000	0.0000
27	0.1572	0.0154	0.0000	0.3351	0.0726	0.0000	0.1538	0.0769	0.0000	0.2000	0.1000	0.0000
28	0.0314	0.0002	0.0000	0.3162	0.0166	0.0000	0.1038	0.0142	0.0000	0.3000	0.0000	0.0000
29	0.1488	0.0327	0.0006	0.5413	0.1986	0.0168	0.4068	0.1695	0.0085	0.8000	0.4000	0.0000
30	0.0578	0.0017	0.0000	0.1511	0.0201	0.0000	0.2250	0.0250	0.0000	0.2000	0.0000	0.0000

Table 8: Per-topic results for 42 automatic runs using the note topics.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.0687	0.0005	0.0000	0.2748	0.0185	0.0000	0.2656	0.0156	0.0000	0.4000	0.0000	0.0000
2	0.0711	0.0038	0.0000	0.1868	0.0423	0.0000	0.2353	0.0588	0.0000	0.5000	0.0000	0.0000
3	0.0099	0.0019	0.0000	0.1745	0.0480	0.0000	0.1118	0.0395	0.0000	0.4000	0.0000	0.0000
4	0.0043	0.0003	0.0000	0.0227	0.0094	0.0000	0.0556	0.0000	0.0000	0.1000	0.0000	0.0000
5	0.0305	0.0000	0.0000	0.1830	0.0000	0.0000	0.1327	0.0000	0.0000	0.5000	0.0000	0.0000
6	0.0485	0.0220	0.0004	0.2666	0.1783	0.0306	0.2482	0.1560	0.0071	0.7000	0.3000	0.0000
7	0.0256	0.0024	0.0000	0.1336	0.0417	0.0000	0.1558	0.0390	0.0000	0.4000	0.0000	0.0000
8	0.0344	0.0238	0.0079	0.7580	0.4229	0.2573	0.3091	0.1311	0.0234	1.0000	0.8000	0.3000
9	0.0196	0.0055	0.0000	0.1758	0.0648	0.0000	0.1382	0.0813	0.0000	0.4000	0.1000	0.0000
10	0.0650	0.0000	0.0000	0.2227	0.0000	0.0000	0.2632	0.0000	0.0000	0.1000	0.0000	0.0000
11	0.0464	0.0115	0.0000	0.4815	0.2160	0.0000	0.2262	0.1553	0.0000	0.9000	0.4000	0.0000
12	0.0345	0.0107	0.0000	0.2953	0.1247	0.0000	0.2252	0.1081	0.0000	0.5000	0.2000	0.0000
13	0.0145	0.0031	0.0000	0.1737	0.0513	0.0000	0.1382	0.0395	0.0000	0.4000	0.1000	0.0000
14	0.0240	0.0053	0.0000	0.2107	0.0718	0.0000	0.1083	0.0583	0.0000	0.3000	0.1000	0.0000
15	0.0225	0.0028	0.0000	0.1615	0.0463	0.0000	0.1558	0.0519	0.0000	0.4000	0.2000	0.0000
16	0.0055	0.0000	0.0000	0.0831	0.0077	0.0000	0.0815	0.0109	0.0000	0.3000	0.0000	0.0000
17	0.0070	0.0016	0.0000	0.1137	0.0351	0.0000	0.1017	0.0395	0.0000	0.2000	0.1000	0.0000
18	0.0236	0.0039	0.0000	0.1836	0.0446	0.0000	0.1324	0.0588	0.0000	0.4000	0.1000	0.0000
19	0.0222	0.0044	0.0000	0.2034	0.1026	0.0000	0.1351	0.0676	0.0000	0.6000	0.1000	0.0000
20	0.0561	0.0313	0.0005	0.8930	0.5723	0.0222	0.3678	0.1809	0.0091	1.0000	0.8000	0.2000
21	0.0966	0.0015	0.0000	0.3065	0.0183	0.0000	0.2466	0.0274	0.0000	0.4000	0.1000	0.0000
22	0.0026	0.0000	0.0000	0.0292	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	0.0119	0.0016	0.0000	0.1175	0.0366	0.0000	0.0917	0.0183	0.0000	0.3000	0.0000	0.0000
24	0.0413	0.0168	0.0019	0.6440	0.4439	0.1340	0.2887	0.1495	0.0168	0.9000	0.6000	0.2000
25	0.0466	0.0055	0.0000	0.4341	0.1075	0.0000	0.2511	0.1050	0.0000	0.7000	0.1000	0.0000
26	0.1161	0.0121	0.0000	0.4175	0.1278	0.0000	0.3009	0.1062	0.0000	0.8000	0.2000	0.0000
27	0.1118	0.0080	0.0000	0.2996	0.0787	0.0000	0.1538	0.0769	0.0000	0.2000	0.0000	0.0000
28	0.0406	0.0029	0.0000	0.3917	0.0821	0.0000	0.2217	0.0425	0.0000	0.7000	0.1000	0.0000
29	0.0342	0.0086	0.0000	0.2223	0.1190	0.0000	0.2119	0.1017	0.0000	0.5000	0.1000	0.0000
30	0.0572	0.0025	0.0000	0.1917	0.0168	0.0000	0.2250	0.0250	0.0000	0.3000	0.1000	0.0000

Table 9: Per-topic results for 19 automatic runs using the description topics.

	infAP			infNDCG			R-prec			P @ 10		
Topic	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst	Best	Median	Worst
1	0.1643	0.0403	0.0028	0.4920	0.2473	0.0481	0.3203	0.2422	0.0312	0.8000	0.3000	0.0000
2	0.0894	0.0114	0.0000	0.3119	0.0875	0.0000	0.2353	0.0882	0.0000	0.4000	0.1000	0.0000
3	0.0974	0.0021	0.0000	0.4817	0.0501	0.0000	0.1842	0.0197	0.0000	0.6000	0.1000	0.0000
4	0.0958	0.0000	0.0000	0.2748	0.0000	0.0000	0.1111	0.0000	0.0000	0.1000	0.0000	0.0000
5	0.0489	0.0084	0.0000	0.3168	0.0793	0.0000	0.1633	0.0714	0.0000	0.6000	0.1000	0.0000
6	0.0572	0.0235	0.0000	0.3210	0.1806	0.0000	0.2837	0.1844	0.0000	0.6000	0.3000	0.0000
7	0.1114	0.0200	0.0000	0.4028	0.1370	0.0000	0.3117	0.1299	0.0000	0.7000	0.3000	0.0000
8	0.0484	0.0345	0.0034	0.7761	0.6589	0.1616	0.3009	0.1979	0.0094	1.0000	0.9000	0.7000
9	0.0832	0.0328	0.0000	0.3775	0.2160	0.0000	0.2846	0.1789	0.0000	0.8000	0.4000	0.0000
10	0.1019	0.0210	0.0000	0.2983	0.1272	0.0000	0.2632	0.0526	0.0000	0.3000	0.0000	0.0000
11	0.0653	0.0223	0.0005	0.5343	0.3174	0.0241	0.2589	0.1853	0.0109	0.8000	0.5000	0.0000
12	0.0462	0.0159	0.0000	0.3391	0.1510	0.0000	0.1982	0.1171	0.0000	0.5000	0.2000	0.0000
13	0.0642	0.0149	0.0000	0.3140	0.1381	0.0000	0.2171	0.1316	0.0000	0.6000	0.3000	0.0000
14	0.1132	0.0085	0.0000	0.5141	0.1040	0.0000	0.2500	0.1000	0.0000	0.7000	0.2000	0.0000
15	0.0515	0.0064	0.0000	0.2777	0.0914	0.0000	0.2338	0.1039	0.0000	0.3000	0.1000	0.0000
16	0.0590	0.0049	0.0000	0.4313	0.0939	0.0000	0.1685	0.0543	0.0000	0.8000	0.2000	0.0000
17	0.1006	0.0225	0.0016	0.5215	0.2010	0.0207	0.3333	0.1469	0.0169	0.8000	0.3000	0.0000
18	0.1220	0.0061	0.0000	0.3335	0.0581	0.0000	0.2794	0.0588	0.0000	0.7000	0.1000	0.0000
19	0.0547	0.0046	0.0002	0.3634	0.0983	0.0100	0.1892	0.0811	0.0045	0.8000	0.1000	0.0000
20	0.0577	0.0452	0.0033	0.8670	0.7163	0.1284	0.3617	0.2857	0.0152	1.0000	0.8000	0.3000
21	0.0161	0.0033	0.0000	0.1662	0.0426	0.0000	0.1233	0.0548	0.0000	0.2000	0.1000	0.0000
22	0.0892	0.0085	0.0000	0.3619	0.0637	0.0000	0.3750	0.1250	0.0000	0.3000	0.1000	0.0000
23	0.0328	0.0010	0.0000	0.3530	0.0370	0.0000	0.1284	0.0367	0.0000	0.4000	0.0000	0.0000
24	0.0378	0.0244	0.0010	0.6631	0.5139	0.0409	0.3067	0.1753	0.0077	1.0000	0.7000	0.1000
25	0.0881	0.0094	0.0000	0.6142	0.1459	0.0036	0.2831	0.0868	0.0137	0.7000	0.2000	0.0000
26	0.1480	0.0439	0.0000	0.4808	0.2632	0.0000	0.3540	0.2389	0.0000	0.9000	0.4000	0.0000
27	0.0716	0.0071	0.0000	0.2183	0.0751	0.0000	0.1538	0.0000	0.0000	0.2000	0.0000	0.0000
28	0.1034	0.0067	0.0000	0.5830	0.1229	0.0000	0.2642	0.0566	0.0047	0.7000	0.2000	0.0000
29	0.2505	0.1069	0.0007	0.7152	0.4161	0.0105	0.4237	0.3051	0.0085	1.0000	0.7000	0.0000
30	0.1357	0.0312	0.0003	0.4256	0.1429	0.0078	0.3000	0.1500	0.0000	0.6000	0.2000	0.0000

Table 10: Per-topic results for 46 automatic runs using the summary topics.

Figure 1: Top manual results (infNDCG). Only shows best run for each participant.

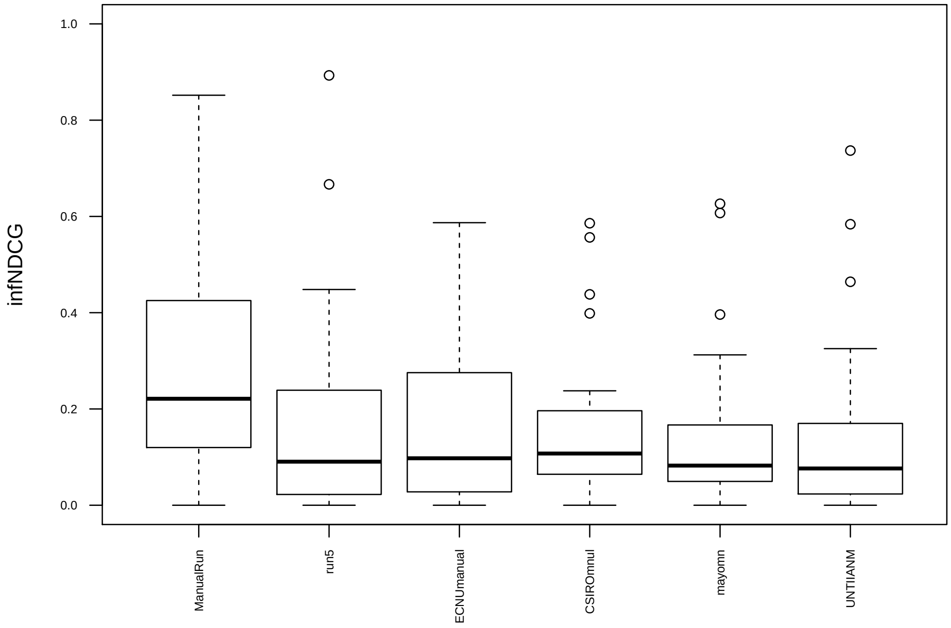


Figure 2: Top manual results (P@10). Only shows best run for each participant.

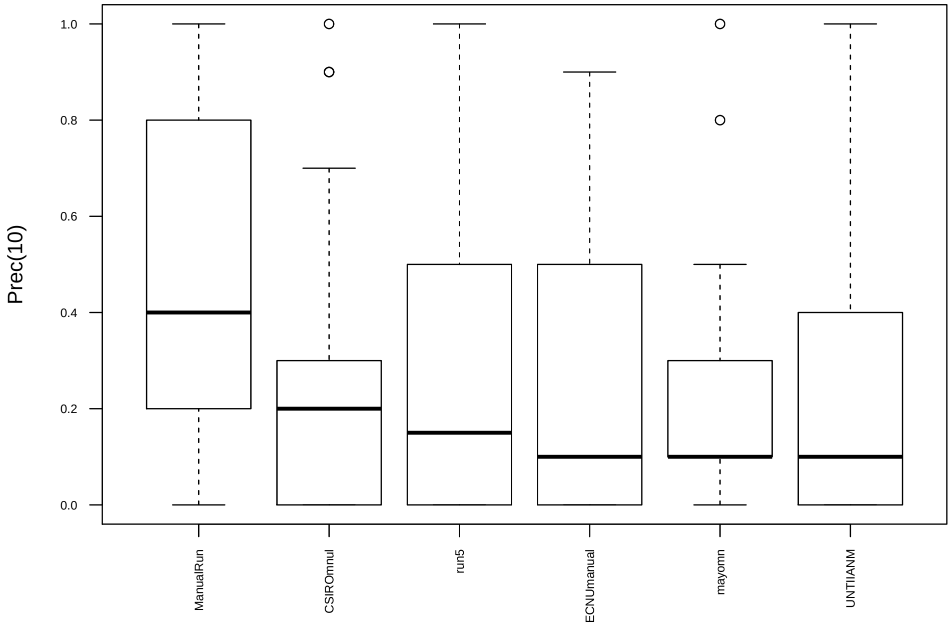


Figure 3: Top automatic results for note topics (infNDCG). Only shows best run for each participant.

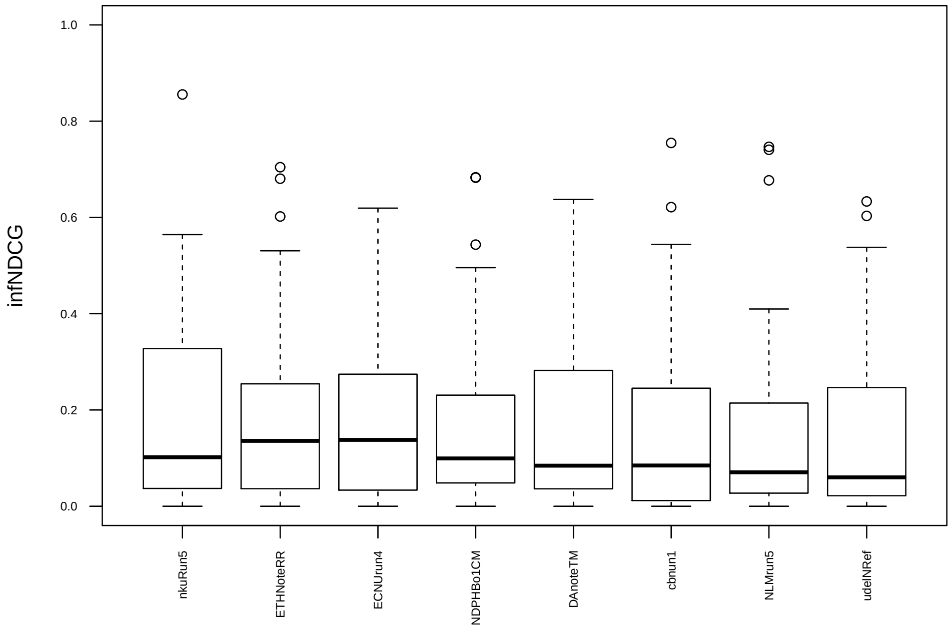


Figure 4: Top automatic results for note topics (P@10). Only shows best run for each participant.

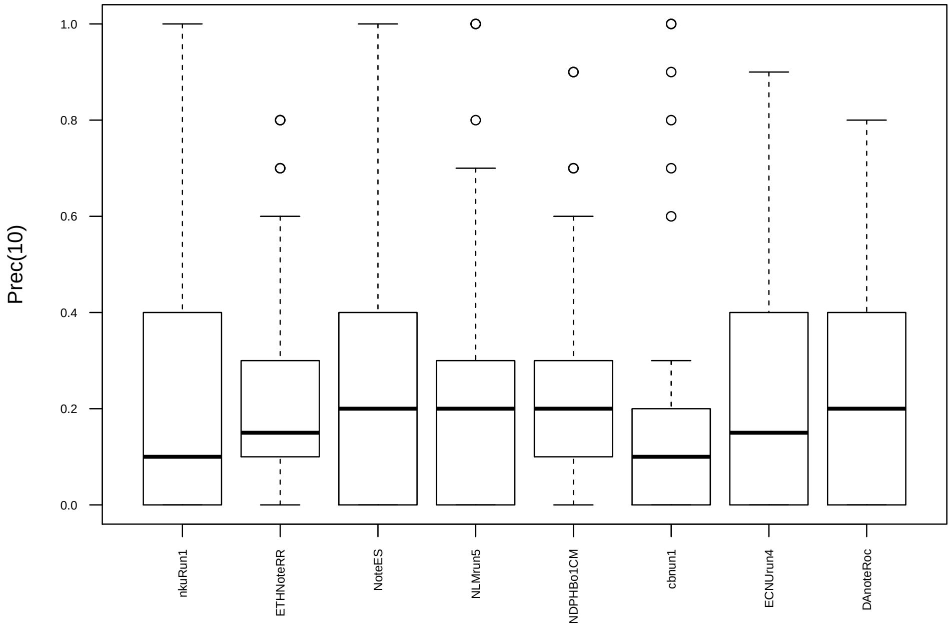


Figure 5: Top automatic results for description topics (infNDCG). Only shows best run for each participant.

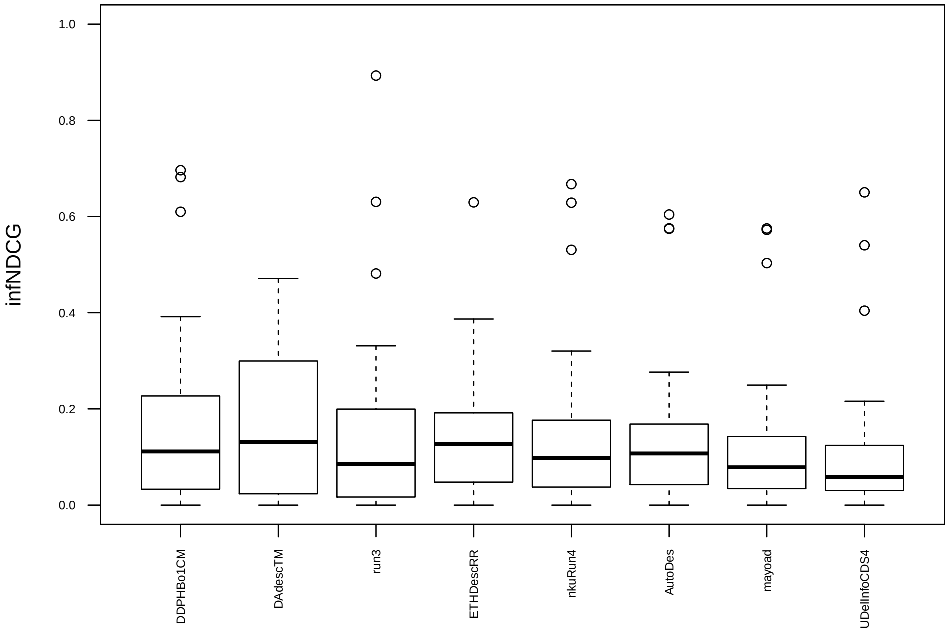


Figure 6: Top automatic results for description topics (P@10). Only shows best run for each participant.

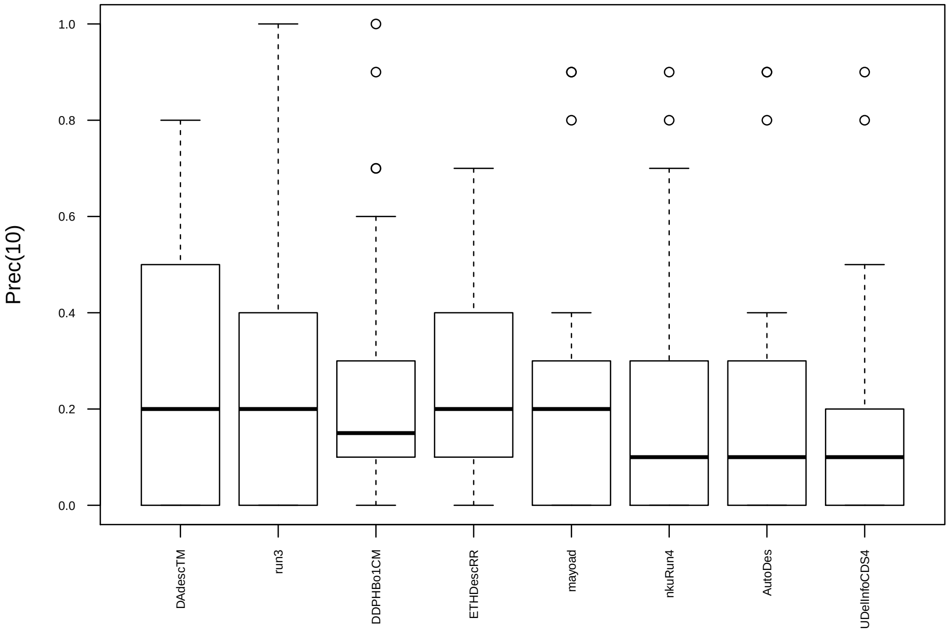


Figure 7: Top automatic results for summary topics (infNDCG). Only shows best run for each participant.

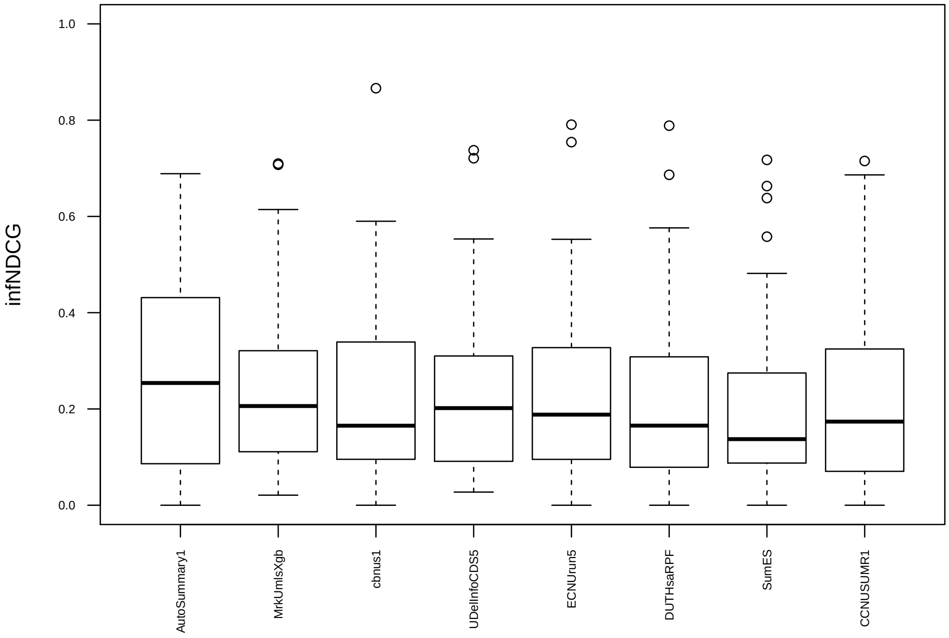
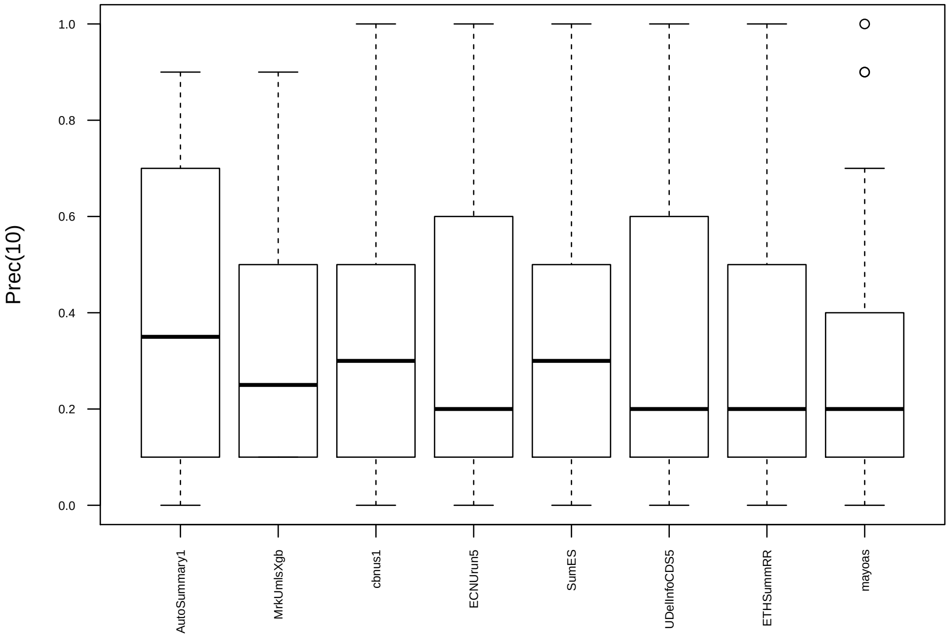


Figure 8: Top automatic results for summary topics (P@10). Only shows best run for each participant.


Prec(10)
AutoSummary1
MrkUmlsXgb
cbnus1
ECNUrun5
SumES
UDelInfoCDS5
ETHSummRR
mayoas