TREC CAsT 2021: The Conversational Assistance Track Overview

Jeffrey Dalton[1], Chenyan Xiong[2], and Jamie Callan[3]
University of Glasgow[1], Microsoft Research[2], Carnegie Mellon University[3]
jeff.dalton@glasgow.ac.uk[1], chenyan.xiong@microsoft.com[2], callan@cs.cmu.edu[3]

1 INTRODUCTION

CAsT 2021 is the third year of the Conversational Assistance Track. The techniques for conversational search continue to evolve as the task becomes more challenging. Proven neural query rewriting and ranking approaches based on pre-trained language models continue to improve with new large-scale datasets. As there is increased dependence on long result history, models that discriminatively select relevant parts of the conversation history are increasingly important. The traditional NLP approaches continue to be used, but generative approaches based on large-scale pre-trained language models are most widely used. One important development this year is the use of dense retrieval approaches. The results show that these models are complementary to traditional search approaches and appear to improve recall, but still usually require a multi-pass neural re-ranking model to be most effective.

Based on participant feedback, CAsT 2021 task is similar to previous years. The task is to identify relevant content for conversational queries that evolve through a trajectory of a discussion on a topic. For 2021, the collection evolved to be based on documents rather than passages to facilitate more complex types of discourse. The collection is similar to previous years and includes MS MARCO documents, an updated dump of Wikipedia from the KILT benchmark, and the Washington Post V4 collection. The collection content is similar to previous years, but with recent content from Wikipedia and WaPo included.

One important change for CAsT 2021 is that every turn has a single manually selected *canonical response* passage result representing a previous system response. This evolved from last year when only some turns were manually selected and others automatically added from baselines. This year's manual results provide consistency between automatic and manual runs. The canonical results are used more, with greater query dependence on previous system responses.

Another minor change to make the conversations more realistic compared with previous years is that the turns introduce simple forms of user revealment, reformulation, and explicit feedback if the previous canonical response is not relevant. This makes the task a bit more realistic by having varying types of user interactions.

Similar to previous years, the topics in 2021 are based on real user needs from information-seeking sessions in Bing sessions [6]. The organizers manually reviewed and filtered sessions to ensure they have meaningful trajectories that are then manually rewritten to make them conversational. The topics reflect diverse types of exploratory information needs while also being grounded in real information needs that have content available in the target collection. We detail topic construction in Section 2.

Table 1: CAsT 2021 Topic 112.

Title: Steroid use in US sports
Description: The history of steroid use in US sports.

Turn	Conversation Utterances
1	What's the history of steroid use in sports in the US?
2	What were Ziegler's improvements?
3	Why are they banned?
4	Are there visible signs?
5	That sounds easy to spot. How do they get away with it?
6	What is the NFL policy?
7	Isn't that speed?
8	What is the difference between the two policies?
9	I heard it even affects card players. Didn't bridge also have a problem?
10	I know what bridge is. I heard there was a drug scandal recently.
11	Does the article have more about it?

Year three continued to have strong participation from more than a dozen teams worldwide. There remains a large gap in effectiveness between manual and automatic systems indicating that there is significant headroom for improving query understanding. In particular, turns with long-distance dependence across multiple turns and queries are very challenging.

We see CAsT continue to evolve as systems become more capable. This year presented a shift towards retrieving fixed passages in the context of a document. We envision conversations that are more natural with the use of initiative to support more varied types of results and more realistic discourse for complex information tasks.

2 TASK, DATA, AND RESOURCES

The core of the CAsT 2021 task remains mostly unchanged from previous years. The goal of the task is to satisfy a user's complex information need expressed through multi-turn conversational queries/utterances (u) for each turn $T = \{u_1, ...u_i...u_n\}$. The change is that these passages come from a document corpus by retrieving and ranking documents and passages from MS MARCO [1], Wikipedia – the KILT dump [5], and news from the Washington Post V4 collection. Previous CAsT overview papers detail the previous versions of the passage collections. [3, 4].

CAsT 2021 has 26 information needs (topics) with an average length of 9.2 utterances, for a total of 239 turns. In comparison, the CAsT 2020 topics are slightly shorter with an average of 8.6 utterances per topic. An example of a 2021 topic is shown in Table 1.

The rest of this section focuses on the major changes for the third year: the more diverse types of interactions and the increased dependence on previous system responses.

TREC'21, November 2021, Virtual

Information Needs. The high-level method for constructing and filtering topics remains the same. Information needs are based on long sessions from a commercial search engine. Once sessions are filtered, the organizer interacts with a baseline CAsT system that includes rewriting and neural re-ranking (made available to participants) and selects a canonical response passage. This is selected to create challenging conversational trajectories. As a result, each turn has a manually selected passage that can be referred to later in the conversation.

Collection. Similar to previous years, the collection includes MS MARCO documents, an updated dump of Wikipedia from the KILT benchmark, and the Washington Post V4 collection.

Due to the presence of duplicates within and across the corpora we identify similar documents using SimHashing [2] with a 64-bit hash. After initial experimentation, we set the duplicate threshold to be less than 5 bits. We post-process the identified duplicates to remove false positives based on a Jaccard similarity greater than 0.85. The result are duplicates that are excluded from the collection / assessment.

Furthermore, to facilitate passage retrieval from the document corpus, we split each document into passages of at most 250 words using version 3.0.6 of the spaCy toolkit with the en_core_web_sm-3.0.0 model. We provide participants with the duplicates list for their de-duplication efforts and tools for passage segmentation.

Interactive CAsT (iCAsT) System. The baseline interactive system[1] features a web interface built on top of the well-performing re-writing and retrieval systems from previous years. The system re-writes queries with a T5-based query re-writer fine-tuned on the CANARD dataset[2]. This uses all previous turn queries and the three previous turn canonical passage responses as context, subject to length constraints. For turns with fewer than three previous turns, the re-writer uses all the available previous turn queries and canonical responses (none for the first turn). When context becomes too large (i.e more than the 512 tokens), the re-writer truncates content from the previous turn passages and queries (oldest first) to accommodate context.

For search, the system retrieves an initial 50 (for interaction) candidate documents using BM25 from the collection. It then segments candidate documents into passages for re-ranking based on user-specified parameters from the web interface. Passage are re-ranked using a T5 based passage re-ranker trained on MS MARCO and available through Pygaggle[3]. The passages are clustered by document in order of max passage score.

Canonical Response. Introduced last year, the interactive nature of the response dependence requires including a fixed system response as part of the benchmark. A key difference is that every turn has a manually selected passage. In year two this was only a subset of turns.

In year three the organizers provide a single *canonical passage response* manually selected for all turns. The response is often selected from results in the baseline results that the organizer finds engaging. The organizers use the baseline system to check the number and nature of answer passages and also ensure challenging conversational structure. The canonical results are usually at least partially relevant, although when the baseline fails the organizer decides to select an irrelevant result or a passage from a different source to advance the conversation. There is only one canonical response for both manual and automatic queries.

Interactive topic development with manual canonical results required significant time commitment from the organizers. It took approximately six hours per topic to iteratively develop and refine the turn trajectories.

2.1 Result Dependence Considerations

The introduction of response dependence more closely models the challenges faced in conversational search systems, but the offline nature of this benchmark and our goal to ensure re-usability introduces challenges and limitations.

The issue remains that there is only one response provided. Because there is often dependence on the result as a building block for later turns, there is a greater need for the result to be at least partially related to the conversational trajectory. As a result, the canonical responses may not truly reflect the quality of an automatic system.

There are three categories in this year's runs based on the data used in the *testing phrase*:

1. *Manual:* Runs that use the manually rewritten (resolved) context-free queries, and/or manual canonical responses.
2. *Automatic-Canonical:* Automatic runs that use the provided automatic canonical system responses.
3. *Automatic-Raw:* Runs that only use the provided raw conversational queries, automatic baseline results, automatic query re-writes, and/or other data sources that do only contain manual or automatic-canonical information.

This year all Automatic runs are reported together. The increased level of result dependence means that automatic systems that don't use the canonical response (Raw) are at a significant disadvantage by ignoring previous conversation context.

Feedback and Revealment In combination with fixed canonical responses, this year also introduced feedback and revealment discourse types. These were added by the organizers and not present in the original search sessions.

For feedback, this year includes turns with explicit relevance feedback on the canonical result. The results from the baseline system may or may not be relevant to the previous user request. In these cases, the organizer had a choice to introduce a feedback turn ("What? No, I want to know..") to give a hint to the system, carry on from the result provided ("No, I meant the funny car. But, that's interesting..."), or change topics ("That's not what I wanted. How about recent developments.."). In all cases, the feedback or reformulation was explicit and fully contained in a single turn.

This year also introduced simplistic forms of user revealment. In some (rare) cases, the simulated user revealed information that is part of a turn, "I live in Seattle and have a big lawn." or "I'm a runner and I've been feeling tired.", or "I'd like a more scientific explanation.". Some of these elements were required for subsequent turns in the conversation.

[1] https://github.com/grill-lab/Interactive-CAsT
[2] https://huggingface.co/castorini/t5-base-canard
[3] https://github.com/castorini/pygaggle

Table 2: Participants and their runs.

Group	Run ID	Run Type	Group	Run ID	Run Type
CFDA_CLIP	CFDA_CLIP_ARUN1	canonical	MLIA-LIP6	t5colbert	canonical
CFDA_CLIP	CFDA_CLIP_ARUN2	canonical	RUIR	RUIR1_TURN-FT	manual
CFDA_CLIP	CFDA_CLIP_MRUN1	manual	RUIR	RUIR2_TURN	manual
CFDA_CLIP	CFDA_CLIP_MRUN2	manual	RUIR	RUIR4_HIST	manual
CMU-LTI	LTI-entity-g	manual	TKB48	bm25_automatic	raw
CMU-LTI	LTI-rewriter-5q	canonical	TKB48	dense_manual	manual
CMU-LTI	LTI-rewriter-g	canonical	TKB48	hybrid_manual	manual
CMU-LTI	LTI-rewriter-tc	canonical	TKB48	sparse_manual	manual
CNR	CNR-run1	raw	UAmsterdam	astypalaia256	canonical
CNR	CNR-run2	canonical	UAmsterdam	historyonly	raw
CNR	CNR-run3	raw	UAmsterdam	historyonlyKILT	raw
CNR	CNR-run4	canonical	UiS	UiS_raft	manual
h2oloo	cqe	canonical	UMD	umd2021_run1	canonical
h2oloo	cqe-t5	canonical	UMD	umd2021_run2doc	canonical
h2oloo	mono-duo-rerank	canonical	UMD	umd2021_run3rrf	canonical
h2oloo	t5	canonical	UMD	umd2021_run4den	canonical
HBKU	HBKU_CQR_POS	canonical	uogTr	uogTrADT	raw
HBKU	HBKU_CQR_TC	canonical	uogTr	uogTrMDT	manual
HBKU	HBKU_CQR-HC	raw	uogTr	uogTrTCT	canonical
HBKU	HBKU_CQRHC_BM25	canonical	uogTr	uogTrTDT	canonical
IITD-DBAI	IITD-RAW_U_T5_1	raw	V-Ryerson	DPH-auto-rye	canonical
IITD-DBAI	IITD-RAW_U_T5_2	raw	V-Ryerson	DPH-manual-rye	manual
MLIA-LIP6	Rewritt5_monot5	canonical	WaterlooClarke	clarke-auto	raw
MLIA-LIP6	t5_doc2query	canonical	WaterlooClarke	clarke-cc	canonical
MLIA-LIP6	t5_monot5	canonical	WaterlooClarke	clarke-manual	manual

Generated Baseline Runs. The organizers generate seven baseline runs this year including ones that use traditional sparse retrieval and those using dense retrieval. These are described in detail below:

(1) ***org_auto_bm25_t5***: Based on the baseline interactive system, this uses the T5-based query re-writer fine-tuned on the CANARD dataset for generative query rewriting on the raw utterances. As with the interactive system, the re-writer uses all previous turn queries and the three previous canonical passage responses as context. In addition, where turns had fewer than three previous turns (e.g turn 1), the re-writer uses all the available previous turn queries and canonical responses. No extra considerations are made for instances where the context might have been too large (e.g deep turns, long paragraphs). For retrieval, BM25 (k1=4.46, b=0.82) is used to collect the top 1000 documents from the collection. These are segmented into sentence-based passages (using spaCy’s SentenceRecongnizer model) with a maximum length of 250 words. The passages are re-ranked with a pointwise (mono) T5 passage ranker trained on MS MARCO. The run file returns the top thousand passages for each query.

(2) ***org_auto_convdr***: This uses ConvDR to encode the conversational query and passages in the collection. It performs Maximum Inner Product Search to retrieve 100 candidate responses. The query encoder is trained on TREC CAsT 2020 data and the encoder uses the concatenation of historical queries, the last system response, and the current query. The passage encoder is from ANCE.

(3) ***org_auto_convdr_bert***: This performs re-ranking of the top 100 passages from convdr. For re-ranking, it uses a BERT-based pointwise ranker pre-trained on MS MARCO and further fine-tuned on CAsT 2020 using joint supervision of manual utterances and relevance labels.

(4) ***org_manual_bm25***: This uses BM25 from Pyserini to perform passage retrieval on the collection with the default configuration and manual utterances for retrieval.

(5) ***org_manual_bm25_t5***: This follows the same setup as *org_auto_bm25_t5* (document retrieval then passage segmentation and re-ranking) using the manually rewritten utterances.

(6) ***org_manual_ance***: This uses an ANCE checkpoint trained on MS MARCO Passages to encode passages and manually rewritten utterances and retrieves the most similar.

(7) ***org_manual_ance_bert***: A fine-tuned BERT-base re-ranks the top 100 retrieved passages from *org_manual_ance*.

Other Data Resources. The organizers released similar resources released as year two [4] including canonical results and baseline runs. In addition, the organizers release the contextual dependence labels and types of queries developed during topic curation. These were created during topic construction and verified, similar to the previous year. We use these annotations to analyze the influence of contextual dependence discussed later.

Evaluation. The overlap among submissions was high again this year. The evaluation was similar to last year, except that judgments were performed at the document level, as described below.

The organizers opted to build shallow pools from a larger number of topics. The original pools were formed using the top ten *passages* from up to four runs per group (a total of 57 runs). Based on assessing speeds last year, it seemed likely that it would be possible to fully assess at least 20 topics. The initial plan was to assess all queries for all topics.

Just before assessing, NIST discovered that some runs contained invalid passage identifiers. A brief investigation determined that different versions or configurations of spaCy produced differing segmentations of documents. Seven groups had at least one run with invalid passage ids. As a result, NIST converted the passage-oriented pools to document-oriented pools. Passage identifiers are truncated to remove the passage id and a max passage algorithm was used to convert passage runs to document runs. Duplicate retrieved documents are removed from the rankings. The top seven documents from each run are used to construct pools from 50 participant runs and 7 organizer runs.

Because assessing documents is slower than passages, pools are shallower and fewer topics/turns assessed. The final qrels were built from depth-7 pools across all submissions (plus the seven baseline runs from the organizers). Nineteen topics have at least some turns judged, most of which were judged through turn eight. Some topics had additional turns judged and a few have less than eight turns judged (125, 128). Turn 117_7 was removed due to only one relevant doc. Topics 109, 114, 120, 122, 126, and 130 were not assessed due to resource constraints.

The CAsT organizers thank Ian Soboroff and Ellen Voorhees for responding quickly and gracefully to the unanticipated, last-minute changes that the document segmentation issues introduced.

3 PARTICIPANTS

CAsT received 50 run submissions from 15 teams shown in Table 2. Participants provided metadata and descriptions of their runs.

Similar to last year, many teams used a multi-step pipeline consisting of 1) conversational rewriting (most incorporating the previous canonical responses), 2) retrieval using traditional IR or dense model, and 3) re-ranking with a neural language model fine-tuned for pointwise (mono) and pairwise (duo) ranking. Almost all teams leverage pre-trained Transformer-based language models for rewriting (BART, T5) and ranking (BERT, ALBERT, T5). Some teams also perform document expansion with generated queries (doc2query). Multiple teams use dense retrieval for passage retrieval and experiment with conversationally encoding queries with result context rather than relying solely on generative query rewriting.

4 OVERALL RESULTS

In this section, we present the results of the submitted runs. We include seven organizer baselines (prefixed *org*) described above that are available in the public CAsT Github repository.

The main results are turn-level macro-averaged response effectiveness. We use four standard evaluation measures: Recall, Mean Average Precision (MAP@500), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@500). The primary measure continues to be NDCG@3 to focus on high-precision and quality responses in the top ranks. Note that we use 500 instead of the typical 1000 because of the conversion from passages to documents. Consistent with previous years, we threshold the official results using a relevance cutoff of **two** as positive for binary measures, because the value of one is marginal in the guidelines.

Table 3: Automatic response retrieval results. Evaluation at retrieval cutoff of 500 with a binary relevance threshold of 2.

Group	Run	Recall	MAP	MRR	NDCG	NDCG@3
h2oloo	mono-duo-rerank	0.850	0.376	0.679	0.636	0.526
WaterlooClarke	clarke-cc	0.869	0.362	0.684	0.640	0.514
h2oloo	cqe-t5	0.846	0.342	0.644	0.618	0.488
HBKU	HBKU_CQR_TC	0.696	0.310	0.632	0.540	0.477
HBKU	HBKU_CQRHC_BM25	0.598	0.287	0.622	0.490	0.471
HBKU	HBKU_CQR_POS	0.588	0.283	0.616	0.487	0.451
CFDA_CLIP	CFDA_CLIP_ARUN1	0.697	0.308	0.613	0.539	0.444
CFDA_CLIP	CFDA_CLIP_ARUN2	0.652	0.301	0.608	0.518	0.439
h2oloo	cqe	0.791	0.289	0.603	0.557	0.438
MLIA-LIP6	t5_doc2query	0.761	0.290	0.585	0.548	0.436
–	**org_auto_bm25_t5**	0.636	0.291	0.607	0.504	0.436
UMD	umd2021_run3rrf	0.723	0.298	0.611	0.539	0.425
–	**org_convdr_bert**	0.426	0.236	0.607	0.398	0.423
uogTr	uogTrADT	0.661	0.278	0.581	0.501	0.417
UMD	umd2021_run2doc	0.613	0.262	0.558	0.478	0.399
HBKU	HBKU_CQR-HC	0.531	0.236	0.531	0.422	0.392
UMD	umd2021_run1	0.613	0.250	0.544	0.464	0.389
MLIA-LIP6	t5_monot5	0.360	0.190	0.571	0.337	0.388
IITD-DBAI	IITD-RAW_U_T5_2	0.327	0.175	0.515	0.316	0.380
h2oloo	t5	0.364	0.176	0.534	0.336	0.377
UMD	umd2021_run4den	0.735	0.265	0.521	0.512	0.377
WaterlooClarke	clarke-auto	0.721	0.260	0.524	0.487	0.375
IITD-DBAI	IITD-RAW_U_T5_1	0.312	0.166	0.509	0.303	0.371
MLIA-LIP6	Rewritt5_monot5	0.361	0.184	0.549	0.332	0.370
CMU-LTI	LTI-rewriter-g	0.465	0.209	0.521	0.386	0.369
CMU-LTI	LTI-rewriter-tc	0.465	0.211	0.528	0.387	0.367
–	**org_convdr**	0.426	0.197	0.505	0.372	0.361
CNR	CNR-run3	0.190	0.123	0.472	0.222	0.349
CNR	CNR-run4	0.187	0.116	0.477	0.220	0.333
uogTr	uogTrTDT	0.557	0.216	0.491	0.408	0.332
uogTr	uogTrTCT	0.562	0.214	0.473	0.414	0.323
TKB48	bm25_automatic	0.623	0.173	0.474	0.405	0.317
CNR	CNR-run2	0.167	0.107	0.444	0.202	0.304
CNR	CNR-run1	0.164	0.101	0.406	0.196	0.298
CMU-LTI	LTI-rewriter-5q	0.392	0.158	0.428	0.319	0.296
UAmsterdam	astypalaia256	0.453	0.120	0.364	0.304	0.236
V-Ryerson	DPH-auto-rye	0.624	0.145	0.367	0.360	0.232
UAmsterdam	historyonlyKILT	0.288	0.084	0.314	0.214	0.196
UAmsterdam	historyonly	0.252	0.077	0.317	0.198	0.195
MLIA-LIP6	t5colbert	0.589	0.076	0.270	0.314	0.154

Table 4: Manual retrieval results. These runs used the manually resolved queries and/or manual canonical results. Evaluation at retrieval cutoff of 500 with a binary relevance threshold of 2.

Group	Run	Recall	MAP	MRR	NDCG	NDCG@3
WaterlooClarke	clarke-manual	0.927	0.473	0.793	0.727	0.644
CFDA_CLIP	CFDA_CLIP_MRUN1	0.806	0.434	0.800	0.669	0.628
CFDA_CLIP	CFDA_CLIP_MRUN2	0.863	0.438	0.792	0.687	0.626
–	**org_manual_bm25_t5**	0.796	0.419	0.780	0.649	0.595
uogTr	uogTrMDT	0.831	0.424	0.777	0.665	0.592
UiS	UiS_raft	0.761	0.397	0.746	0.637	0.579
RUIR	RUIR1_TURN-FT	0.796	0.378	0.717	0.618	0.554
RUIR	RUIR2_TURN	0.796	0.390	0.737	0.626	0.554
–	**org_manual_ance**	0.539	0.308	0.727	0.485	0.548
–	**org_manual_ance_bert**	0.539	0.329	0.702	0.498	0.540
RUIR	RUIR4_HIST	0.796	0.325	0.664	0.593	0.493
CMU-LTI	LTI-entity-g	0.492	0.264	0.632	0.434	0.462
TKB48	hybrid_manual	0.710	0.197	0.601	0.467	0.438
TKB48	dense_manual	0.624	0.181	0.580	0.428	0.417
TKB48	sparse_manual	0.747	0.237	0.595	0.510	0.407
–	**org_manual_bm25**	0.471	0.213	0.594	0.400	0.407
V-Ryerson	DPH-manual-rye	0.188	0.074	0.386	0.170	0.252

Automatic run results. Table 3 shows the results for the 37 automatic runs with a median NDCG@3 score of 0.377. The top two runs use a pipeline that includes mono-duo T5 re-ranking after a first pass retrieval. The most successful first pass retrieval pipelines leverage both generative conversational query rewriting and dense retrieval (on its own or in combination with a traditional retrieval). It's noteworthy that cqe and cqe-t5 use only conversational understanding combined with dense retrieval approaches without further re-ranking and outperform the organizer t5 re-ranking baseline due to their improvements in recall effectiveness. All the other top-performing runs use a combination of a mono/duo re-ranking with a neural language model. We observe that while many of the new conversational dense retrieval approaches perform competitively in recall (t5colbert, DPH-auto-rye), most appear to benefit significantly from further re-ranking for high precision in the top ranks. Note that the organizer convdr runs use the top 100 passages.

Manual run results. Table 4 shows the results for the 13 manual runs with a median NDCG@3 value of 0.554. Three runs outperform the organizer bm25_t5 re-ranking. These runs both use forms of query and/or document expansion. The best performing run, *clarke_manual* performs query expansion on an external collection, dense passage retrieval with ANCE, and a mono/duo re-ranking with T5. This approach achieves very high recall (92.7%), indicating that there is headroom in the re-ranking phase even with manually rewritten queries. The runs from CFDA_CLIP both also include document expansion from doc2query.

Overall. The best automatic runs achieve high recall, approximately 85%, with manual runs only 9% more effective than automatic. The improvement in NDCG@3 between the best manual and automatic runs is 22%, greater than in previous years. This indicates that while the automatic methods identify candidates reasonably well the conversational query understanding remains challenging.

4.1 Results by Topic

Figure 1 provides a per-topic analysis comparing the two classes of systems across topics. It uses data from all submitted runs. As with last year, the results show that the topic difficulty varies widely across topics. We observe that there is a large absolute gap in approximately half the topics overall. The hardest topics for automatic systems are 113 and 117. This is also indicated in the structure of Topic 113, which includes multiple feedback turns when the baseline system fails.

4.2 Results by turn depth

In this section, we discuss how systems perform over the course of the conversation and as turn depth increases. Due to the small sample size, turns beyond nine are truncated. Figure 2 shows the average NDCG@3 at each turn depth for the different categories of systems. To focus on strong systems, the figure only shows the data for runs that perform at or above the median NDCG@3.

For the automatic runs, the results show a steady turn-by-turn drop in system effectiveness from turn one to five. Although there is a slight increase between turns five and eight, system effectiveness drops by 49% from the beginning. In contrast, the Manual runs appear relatively consistent across turn depth, with even a slight increase in effectiveness (except turn 9). The most noticeable pattern is the steady degradation in effectiveness up to turn 5 in automatic runs.

Table 5: Dependence results for automatic runs. We report the average across automatic runs median or better.

Dependence	Turns	Auto. NDCG@3
All turns	158	0.433
None	31	0.513
Query	60	0.429
Query (hard)	16	0.440
Result	86	0.393
Result (hard)	17	0.348

Although the gap between automatic and manual systems appears to widen as turn depth increases, the fact that they follow similar behavior may also indicate the discourse structure challenges in the topics.

4.3 Explicit Query and Turn Dependence

In this section, we study the effect of context. In the organizers' dependence annotations 60 queries depend on a query from a previous turn in the conversation. There are 86 queries that depend on results from previous turns. Note that a turn may depend on multiple previous queries, results, or a combination of both.

Figure 3 shows the breakdown of these dependencies by turn depth according to the source turn that is referenced. Compared to year two, there is a less dependence on the first turn query, and increased result dependencies across all turns. The trend of strong dependence on the previous turns continues this year, with the majority of query dependencies being to the immediate previous turn. About a fourth of the query dependencies are *hard*, defined to reference not to the first or immediate preceding turn.

Table 5 shows the results broken down by different types of conversational contexts. Some turns (for example all first turns) do not rely on a previous turn at all. We label this *None* and they represent approximately 20% of turns. Others only depend on previous utterances *Query.* Less than 40% of all turns depend on a previous utterance with some of these being the first turn. We also further split the dependence into a *Query hard* subset (approximately 10% of turns) where there is a dependence on a previous query that is not the first turn and not the immediately preceding turn. The bottom two rows focus on result dependence. The *Result* type has 54% of turns and *Result Hard* has 11% of turns. Similar to queries, the hard variant for results is result dependence beyond the first or immediate preceding turn.

The results show that turns without dependence are the easiest and systems perform the best on this subset. Systems perform around the same on the *Query* subset as they do on all turns. Interestingly, systems perform slightly better on queries with harder context (*Query hard*) than they do on the *Query* subset, which bears further investigation.

As with last year, turns with result dependence perform are harder than query dependence - an 8% relative reduction compared with query dependence. And the hard subset of result dependencies are even more challenging, with a further 11% relative drop in system effectiveness. As with last year, this continues to highlight areas for further research.

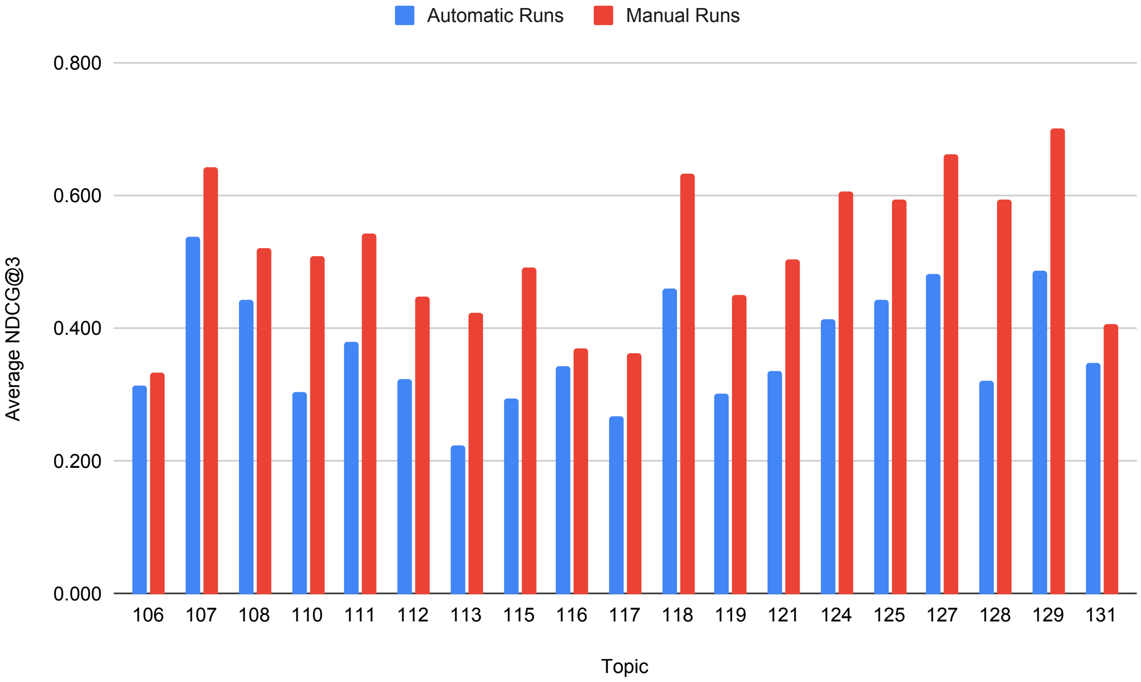

Figure 1: NDCG@3 aggregated for each topic across all runs.

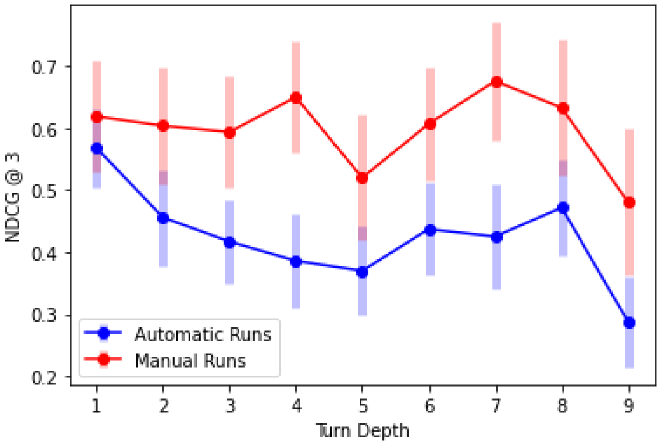

Figure 2: NDCG@3 at varying conversation turn depth. We report the average across runs median or better.

5 CONCLUSION

The third year of TREC CAsT continued developing resources for studying conversational information seeking and added to the community's understanding of the topic. Conversations had more varied types of discourse with feedback turns and elements of user revealment with greater dependence on canonical system response.

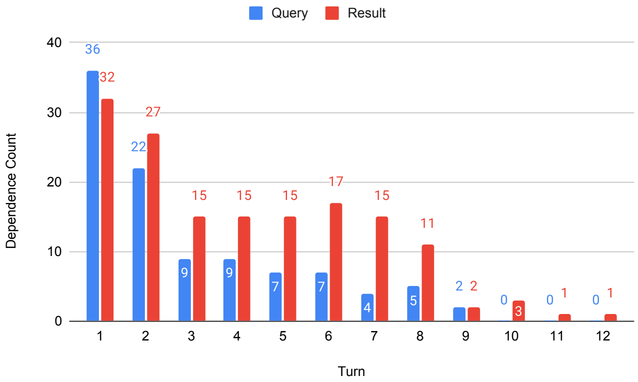

Figure 3: Statistics on the source of contextual information in query and response (used by later turns) by turn depth.

- **Conversational Dense Retrieval.** To overcome issues of increasing sequence length and issues of lexical mismatch the use of dense retrieval for the task was seen for the first time. The results show they can be highly effective for recall but are most effective when combined with neural re-ranking.
- **Conversational Language Understanding.** Sequence-to-sequence generative rewriting methods continue to be widely used and models are improving to handle more complex types of dependence.

- **Conversational Context.** Clean, resolved context remains advantageous for manual runs as the results show that they maintain or even improve in effectiveness over the course of the conversation. In contrast, automatic systems still suffer from degradation in their effectiveness as conversations become longer.
- **Ranking.** The use of pre-trained neural language models for ranking continues to be widely used in the most effective systems. The results show that there is still significant headroom for precision in the early ranks.
- **Conversational structure.** More systems used the canonical responses because this year dependence on system response was very important for effectiveness. The use of the iCAsT interactive system for topic development made canonical responses for all turns possible. We note that injecting canonical responses into systems can lead to artificial behavior. To overcome this, new methods for scalably and interactively developing conversational trajectories on the same topic is an area for the future.

After the success of year three, we look forward to year four.

6 ACKNOWLEDGMENTS

We thank Paul Owoicho for his extensive work with all aspects of the track: interactive baseline system development, topic development, and creation of baseline runs. We thank Shi Yu for contributing the ConvDR dense retrieval baselines and helping with topic construction. We also are deeply thankful for Ellen Voorhees' experience, patience, and persistence in running the assessment process. Finally, we thank all our participants.

REFERENCES

[1] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)

[2] Charikar, M.: Similarity estimation techniques from rounding algorithms. In: STOC '02 (2002)

[3] Dalton, J., Xiong, C., Callan, J.: Cast 2019: The conversational assistance track overview. In: The Twenty-Eighth Text REtrieval Conference Proceedings (TREC 2019). National Institute of Standards and Technology, special publication (2020)

[4] Dalton, J., Xiong, C., Kumar, V., Callan, J.: Cast-19: A dataset for conversational information seeking. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1985–1988. ACM (2020)

[5] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., Cao, N.D., Thorne, J., Jernite, Y., Plachouras, V., Rocktaschel, T., Riedel, S.: Kilt: a benchmark for knowledge intensive language tasks. ArXiv **abs/2009.02252** (2021)

[6] Rosset, C., Xiong, C., Song, X., Campos, D., Craswell, N., Tiwary, S., Bennett, P.: Leading conversational search by suggesting useful questions. In: Proceedings of The Web Conference 2020. pp. 1160–1170 (2020)