TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation

Paul Owoicho[1], Jeffrey Dalton[1], Mohammad Aliannejadi[2], Leif Azzopardi[3], Johanne R. Trippas[4], Svitlana Vakulenko[5]
[1]University of Glasgow, [2]University of Amsterdam, [3]University of Strathclyde, [4]RMIT University, [5]Amazon
p.owoicho.1@research.gla.ac.uk[1], jeff.dalton@glasgow.ac.uk[1],m.aliannejadi@uva.nl[2], leif.azzopardi@strath.ac.uk[3], j.trippas@rmit.edu.au[4], svvakul@amazon.com[5]

1 INTRODUCTION

The fourth year of the TREC Conversational Assistance Track (CAsT) continues to focus on evaluating Conversational Passage Ranking (ConvPR) for information seeking but with several new additions to improve the realism of the task and to improve our understanding of conversational search.

This year the topics were more realistic and dynamic involving branching that created different conversational paths and trajectories through the topic space. These conversational paths motivated introducing and piloting of new evaluation metrics that go beyond independently evaluated turn-based measures to metrics that consider the flow of conversation over the sequences of turns.

Next, the track introduced a response generation sub-task, with retrieved passages used as provenance, to explore the use of generating responses from one or more passages, including both extractive and generative approaches. The response evaluation includes elements of naturalness, conciseness, as well as relevance. The track also included a mixed initiative sub-task, where given the prior conversational context, the task was to generate clarifying or follow-up questions to direct the conversation in a relevant direction. And, for each conversation turn, the system may return a response or ask a question. The system may select one or more questions to ask the user for each turn in a conversation. This also motivated new evaluation measures to be developed.

In sum, the tasks and topics for CAsT were designed to be more challenging and a step closer to a fully conversational system — in line with the developments in core techniques such as conversational query rewriting (CQR), conversational query expansion (CQE), and the continued shift towards the use of dense retrieval and learned sparse representations in combination with hybrid approaches and multi-stage rankers leveraging large pre-trained language models.

The core task is for the system to return a response after every user utterance. Each system response may be a simple passage, but it may also be an extracted or generated summary from one or more passage results. All responses must have at least one passage called 'provenance' from the collection because the primary task evaluation remains passage/provenance ranking. Similar to previous years, the system may use all previous turns in the conversation as context; this is all parents in the conversational topic tree.

To support richer and more realistic conversations, the topic structure changed to be a tree that consists of multiple overlapping conversations on the same topic rather than a simple (or linear) series of user-system turns (in past years). An example topic with a tree structure with parent utterance links is shown in Table 1. The topic tree structure was introduced as different paths evolve based on the results and mixed-initiative that is experienced – and so they provide a greater breadth/coverage of the topic as a result.

Table 1: CAsT 2022 Topic 140.

Title: South America
Description: An exploration of different aspects of South America's culture, tourist attractions, and cuisine.

Turn	Parent	Conversation Utterances
1-1	-	What should I know about Argentina?
1-3	1-2	What makes it the capital?
1-5	1-4	That's not what I meant. I'd like to know about the culture of Buenos Aires.
2-2	2-1	What makes their culture distinct from the rest of the country?
2-4	2-3	What's Merienda?
3-1	2-3	You've mentioned that several times now. Tell me more.
3-3	3-2	I meant the tango.
3-5	3-4	Thanks, but I'm not asking about clothing. Why is it important historically?
4-1	2-1	What are they known for?
5-1	1-2	Mmm, meat sweats. I've heard it's good there, tell me more.
5-3	5-2	No, what are some popular dishes?
—	—	—

Another change from previous CAsT editions is that mixed-initiative responses are included in trajectories. These turns provide the system with a chance to ask the user a question to 1) clarify the information need, 2) ask for feedback, or 3) elicit the task. This new addition aims to make the track more interactive and realistic. This led to the MI sub-task. Participants had the option to submit mixed-initiative (MI) utterances at every point of the conversation and receive a user response. This was organized as a separate phase of the track, but the outcome of this sub-task could be used in the main phase. This represents a first step for the track beyond "*user ask, system reply*", albeit on predefined fixed trajectories.

Mixed initiative is incorporated into the canonical system responses. As a result, some canonical turns have the system ask a question where the user (topic creator) responds appropriately in

TREC'22, November 2022, Virtual

the next turn. This results in groups of related turns that share the same information need. It also provides flexibility to allow some user utterances to be vague, under-specified, etc... For example, starting the conversation with a vaguely defined information need. We provide annotations on these relationships. We only create judgments for and evaluate system effectiveness on the subset of turns that have been clarified to be unambiguous and correctly specified. These innovations allow us to create more natural conversations with richer discourse structure while also retaining reusability.

The use of the new sub-tree structure allows flexibility and more realistic conversations by supporting multiple paths on the same topic. For example, on a trajectory with a relevant result and one with an irrelevant or partially relevant result. It allows conversations of a greater variety of depth and breadth.

Similar to previous years, most 2022 topics are based on real user needs from information-seeking sessions in Bing sessions [6]. The organizers also added a few more diverse topics to expand the themes. The organizers manually reviewed and filtered sessions to ensure they had meaningful trajectories that are manually rewritten to make them conversational. The topics reflect diverse types of exploratory information needs while also being grounded in real information needs with content in the target collection. We detail topic construction in Section 2.1.

Year four continued to have strong participation from more than a dozen teams worldwide. There remains a gap in effectiveness between manual and automatic systems, although this is shrinking, particularly in areas of recall.

We see CAsT continue to evolve as systems become more capable. This year presented a shift towards interaction with user responses such as clarification and feedback being used for the first time. It also presented the first opportunity to test response generation approaches that leverage retrieved passage sources.

2 TASK, DATA, AND RESOURCES

In this section, we describe data and resources used in both the main task as well as the mixed-initiative sub-task. The MI sub-task builds on the main task using the same collection and topics.

2.1 Main Task

A key change to the main task involved how systems returned responses (rather than just returning passage ids). A response is a text suitable for showing to a user. It should be fluent, satisfy their information need, and not contain extraneous or redundant information. A response is limited to a maximum of 250 words (as measured by spaCy v3.3), but should vary depending on an appropriate query-response. We evaluated the quality of the responses using human judgments (described later in this section). The system responses must be grounded in canonical passages from a document corpus consisting of MS MARCO V2 [3], Wikipedia – the KILT dump [5], and news from the Washington Post V4 collection.

For compatibility, the main task remains provenance passage ranking, ConvPR. The final provenance "run" takes the provenance passages for all responses in response order. For compatibility with previous years, the first 1000 provenances for each turn are used. Because a response may have multiple source passages, the score of passages in the provenance list for a response is used to order passages in descending order. If a source passage occurs in multiple responses, it will be ranked by its first response.

CAsT 2022 has 18 information needs (topics) with an average length of 11.39 user utterances with an average of 2.7 sub-topics. There are a total of 205 user utterances, including vague, ambiguous, or user responses to system questions. In comparison, the CAsT 2021 topics are slightly shorter, with an average of 9.2 utterances per topic. A major difference from previous years is that topics in CAsT 2022 follow a "tree" structure with distinct conversational paths. Each topic starts off with a common query, (i.e. "I remember Glasgow hosting COP26 last year, but unfortunately, I was out of the loop. What was it about?") but branches off at various points in the conversation as the topic unfolds. Topics have a maximum of nine distinct conversational paths and a minimum of one. A subset of Topic 140 is shown in Table 1. Note that User and System utterances are decoupled to support branches with multiple different system responses to a single user utterance. The turn structure is encoded by the parent relationship.

Topic Creation. The high-level method for constructing and filtering topics remains the same as in previous years. Information needs are based on long sessions from a commercial search engine. Once sessions are filtered, the organizer interacts with the iCAsT system described in [4], to select a set of passages to synthesize a natural language system response. The system response is a human-written summary of one or more passages. This response takes the place of the canonical passage response from Year 3. The topics include both the human summary as well as passage provenance links that ground the response.

Collection. The current iteration updates the document collection to include MS MARCO V2 documents, keeping the KILT-based Wikipedia dump and Washington Post V4 collections from year 3. This update brings the total number of documents to about 17 million. Due to the size and nature of the collection, we observed that several documents (within and across each collection) contained the same or similar web page hosted on different URLs. We de-duplicated the entire collection by grouping documents from the same website with the same title and cosine similarity greater than 90%. The longest document in each group was treated as the original and included in the final collection. This excluded roughly 1 million documents.

As in year three, we split each document from the de-duplicated collection into canonical passages of at most 250 words using version 3.3.0 of the spaCy toolkit and the en_core_web_sm-3.3.0 model. We provided these canonical passage splits to participants together with python based scripts to allow participants to process the collection themselves as well as MD5 hashes to verify chunking correctness.

Generated Baseline Runs. We generated two baseline runs for the main task this year. For NLU the automatic run - ***BM25_-T5_BART_automatic*** uses a query rewriter trained on CANARD with k-context of both query and results, see the 2021 overview for details. The manual run ***BM25_T5_BART_manual*** uses the manual rewrites provided in the topics file. For both the rest of the processing is fixed. It is a multi-stage pipeline with (1) document retrieval, (2) passage segmentation, and (3) passage reranking with a fine-tuned language model. For document retrieval, we use a BM25 (K1=4.46, b=0.82) from the Pyserini toolkit. The segmenter used the standard spaCy segmentation described above. The first 1000 passages (in document rank order) were re-ranked using a T5 ranker trained on the MS MARCO dataset, available on HuggingFace [1]. The top 1000 ranked passages form the provenance run. For response generation, the baseline generated one response using the top three ranked passages and a standard off-the-shelf abstractive summarizer based on BART[2].

Relevance Assessments The main provenance judgment task was performed by NIST assessors following the scale and methodology of previous years. They only assessed the source passages, not the generated responses. The total pool up to depth 20 contained 49,878 passages, of which 43,027 passages were judged. Only 17 of the 18 topics are judged; topic 134 is not assessed. Additionally, two turns were filtered out (139_2-5 140_4-4) because they did not contain sufficient relevant results (less than 3 results with the relevance of at least 2). Of the total of 205 user turns, there were judgments for 163 user turns. Utterances that were user responses to system questions were not judged. Additionally, vague and unspecified turns were not included, only the clarified versions. This supports turn-level relevance assessment and focuses provenance assessment resources on well-specified turns.

There was an average of 258 judged results per turn, with an average of 76 at least partially relevant. There was a total of 12,318 at least partially relevant results. There are 5053 1s, 3297 2s, 2129 3s, and 1839 4s. Note that, like in previous years, we use a threshold of 2 for binary relevance, with 1 being quite marginal. The relevant distribution by collection is 10,775 MARCO V2, 999 KILT, and 544 WAPO.

Response Quality Judgments. To label the quality of responses, we employed crowd workers. We sourced annotations for relevance, naturalness, and conciseness for the top response across all turns from each submission to the main task. Additionally, we released a mapping of each unique response to a Response ID to support reusable evaluation with the crowd-sourced judgments using any standard IR evaluation toolkit. The response bank contains 2314 unique responses and we released 2479 relevance judgments across all (judged) turns.

We collected these judgments from crowd workers through the Prolific platform. We asked between 5 to 10 workers to assess the responses from the response pools for all turns in one topic. For each turn-response pair, workers see the "conversation so far" as context to make their judgments. The dimensions and rubric for judgments are defined below. Note that the relevance criteria were on a simplified graded scale that differs from those used by the NIST assessors for provenance assessment.

Relevance: Does the response follow on from previous utterances?

- *0. Not Relevant* - The response does not follow the previous utterances; seems to be completely random to the current conversation; seems to be a completely different conversation.
- *1. Partially Relevant* - The response is partially off-topic; may be vaguely related, but too divergent from the conversation.
- *2. Relevant* - The response follows from previous turns, but it is not entirely clear why the response is being presented.
- *3. Highly Relevant* - The response directly follows and it is clear why the response is being presented.

Naturalness: Does the response sound human-like?

- *0. No* - The response does not sound like something a human would say given the conversation.
- *1. Somewhat* - The response is a bit human-like. The response is somewhat understandable but may not be entirely fluent and natural.
- *2. Yes (but not completely)* - The response is almost human-like. The response is well-formed but is not natural.
- *3. Yes* - The response is very human-like and fluent.

Conciseness: Does the response adequately follow the previous utterances in a concise manner?

- *0. No* - The response is too wordy or too short. The response may also contain lots of irrelevant content or no relevant information at all.
- *1. Somewhat* - The response is a bit wordy and does not adequately address the user's utterance (i.e the response is longer than needed).
- *2: Yes (but not completely)* - The response is brief but not comprehensive (i.e does not adequately address the user's utterance/query or properly follow on from the conversation).
- *3: Yes* - The response is brief but comprehensive (the response was concise and to the point without too much/little other information).

To eliminate low-quality workers, we screened the responses for clearly low-quality responses for each topic — where the judgments should be 0 or 1 for each question. These were then used as a check to filter out workers who marked such responses differently from what was expected. Specifically, where a response should be labeled 0 or 1 for relevance, but receives a 3 from a worker then all the worker's judgments for that topic are discarded.

Based on these crowd-sourced judgments, we released files containing the response pool, response bank, and annotations for the three criteria. We determined the final judgment for each response based on a majority vote among the crowd workers. Where there is no majority, the final judgment is the average across all workers. Note that despite the quality filtering process described earlier, each topic had a minimum of three judgments.

2.2 MI Sub-task

Teams are able to test their systems in terms of the generation or selection of MI utterances. At each point in the conversation, a system could pose a clarifying question and receive an answer from a human annotator. We provided a question bank to the teams from which they could select the MI utterance. Given the human-in-the-loop nature of this sub-task, the teams could also opt for a generative model. We collected the MI sub-task submissions a few days prior to the deadline of the main task and crowd-source the responses to the top-ranked utterance of all the submissions. Each team's responses, together with baseline responses were then communicated to the teams one week before the main deadline.

[1] https://huggingface.co/castorini/monot5-base-msmarco
[2] https://huggingface.co/facebook/bart-large-cnn?text

Teams could submit runs to the main task, using the MI utterances and responses.

Question Bank. Following [1, 2], we collected a question bank on the topics in the collection. The idea was to have a set of human-generated questions on each topic, covering various aspects of the topic. To this end, we set up a crowd-sourcing task where we provided the workers with the topic description and asked them to input the query into a commercial search engine of their choice. We then instructed them to scan the first two pages of the results page, as well as query auto-complete and suggestions to get an idea of the different aspects or facets of the query. Finally, we asked each worker to input 6 questions per topic, each focusing on a different aspect. The released question bank contains 4,496 questions.

Crowd-sourced MI Responses. We designed a consequent crowd-sourcing task to collect the user responses to the MI utterances submitted by the teams. In this task, we provided the crowd-workers with the conversation context up to the point where the MI utterance was posed and instructed workers to respond to the MI utterance as if they are a part of the conversation. To give the worker more information about the topic and the user's information needs, we also revealed three turns of the conversation in the future. Therefore, the worker uses some pieces of information that could have been revealed in the future to answer the current question. The intuition behind this decision is based on the fact that MI utterances can be used to clarify or elicit different aspects of a topic that could be naturally revealed as the conversation progresses in a non-mixed-initiative manner. We only revealed three future turns to mimic the situation where an information need evolves within a conversation as the user and system interact.

Baselines. We released 4 baseline systems for the MI Sub-task that used various ranking and generative methods. These are described below:

- **miniLM-bert-mi**: This is a two-step system that uses the distilled miniLM model to generate a candidate set of questions from the question bank (using sentence similarity), and re-ranked them with a BERT model fine-tuned for pair-wise ranking on the QULAC dataset. We performed training and inference with the SentenceTransformers library.
- **bm25-baseline**: This system used a BM25 function to retrieve candidate questions from the question bank.
- **T5-***: The T5-based systems were trained on the ClariQ dataset for clarification question generation. This included *T5-raw* variant that used the raw utterance at each turn for generation and the *T5-rewrite* variant that used the automatic rewritten utterance as input.

Clarification Question Judgments. Following the same methodology for collecting response judgments, we collected clarification question judgments against the following dimensions and criteria:

Relevance: Does the question (logically) follow on from previous utterances?

- *0. Not Relevant* - The question does not follow on from the previous utterances; seems to be completely random, to the current conversation; seems to be a completely different conversation.
- *1. Partially Relevant* - The question veers off-topic; is vaguely related, but too divergent from the conversation.
- *2. Relevant* - The question follows, but it is not entirely clear why the question is being presented.
- *3. Highly Relevant* - The question directly follows, and it is clear why the question is being presented.

Novelty: Does the question add new information to the conversation?

- *0. No* - The question restates the user query; asks a question for which the answer can already be determined from the conversation thus far; restates something already said.
- *1. Somewhat, but no (non-relevant/nonsensical)* - The question adds something new, but it does not make sense in the current conversation.
- *2. Yes (but not useful)* - The question adds something new but is not helpful or interesting to the conversation.
- *3. Yes (adds to the conversation/interest)* - The question adds something new to the conversation that could be interesting to follow up on, or presents paths that could be taken later in the conversation.

Diversity: Does the question provide a number of options?

- *0. No* - The question provides an answer without explicitly trying to provide new avenues for the user to inquire about.
- *1. Offers binary choice (did you mean...)* - The question presents a binary choice, i.e. yes/no or A and B
- *2. Offers 3 or more* - The question offers the user a number of choices on how to proceed.
- *3. Open-ended* - The question invites any number of responses/answers from the user.

We generated and released files containing the question pool, an updated question bank, and relevance judgments for each criterion. The updated question bank contains a 2024 set of judgments, while the question bank contains 5596 unique questions.

3 EVALUATION

This year, we explore new evaluation methodologies for the main task and the MI sub-task. These are discussed in the following section:

3.1 Main Task

Turn Based Evaluation: For the main task, we evaluated the runs across two dimensions given the ranking for each topic turn (i) the ranking depth, and (ii) the turn depth. For ranking depth, we focus on earlier positions (1, 3, 5) for the conversational scenario (where we assume that the top k results will be used to formulate the response back to the user). The turn depth evaluates the run performance at the n-th conversational turn. Performing well on deeper rounds indicates a better ability to understand the preceding context. We use the mean NDCG@3 as the main evaluation metric, with all conversational turns averaged using uniform weights. We also measure the turned-depth measure based on NDCG@3&n, with the per query NDCG@3 scores averaged at depth (n). In addition to the NDCG, we also calculated the precision, recall, Average Precision, and Reciprocal Rank, where again we averaged over all turns.

Conversational Path Evaluation: To evaluate the quality of the overall conversation, we developed a series of new metrics that aimed to summarize the conversational utility given the flow of relevant responses. To do so, we defined three related metrics, that build upon the turn-based evaluation. Given the conversational tree for each topic, we first extract the conversational paths (and so one topic produces many paths). Each path p consists of a series of turns (t). The quality of the response (based on the ranking) given each turn t was evaluated using the standard metrics reported earlier. Given the scores s_t for each turn t, we then computed the overall score for the path p.

The **Conversational Cumulative Gain** (CCG) for a path p is:

$$CCG(s) = \frac{1}{|p|} \sum_{t \in p} s_t \tag{1}$$

where we take the average of the scores s_t over the path (This is analogous to session-based cumulative gain). The CCG score for each path is then averaged over all paths to get the average CCG score for the system. Note, s denotes the underlying metric used to compute the turn level score (i.e., NDCG@3).

The CCG metric treats each turn independently, so the dependence between turns is only weakly captured by averaging over the path. To address this limitation, we developed two additional metrics that tried to encode the dependence turns and flow of the conversation into the metric score. First, we developed a **Conversational Path Score** (CPS) where the score was computed based on the number of consecutive sequences of responses with a score s_t greater than a threshold θ. To compute the CPS, we first split the path into a number of relevant, $r(t_i, ..., t_{i+j})$ and non-relevant sequences, $n(t_i)$, where sequences have a length $|r(.)|$ based on the number of consequence turns where $s_t > \theta$ (i.e., they are considered reasonably relevant to enable the conversation to flow). Then for each relevant sequence, we took the length of the sequence and raised it to an exponent γ, before summing over all relevant sequences. The maximum relevant sequence score would be the total path length raised to the exponent (i.e. $|p|^\gamma$). This was used to normalize the score. Thus the CPS can be formulated as follows:

$$CPS(s) = \frac{1}{|p|^\gamma} \sum_{r(.) \in p} |r(.)|^\gamma \tag{2}$$

where γ defines how much we value the flow of conversation. If $\gamma > 1$ then it implies that longer relevant sequences are considered better and users would receive more gain from the conversation "*flowing*". While if $\gamma < 1$ then it implies that longer relevant sequences are considered worse and users would receive less gain from such "*flowing*" conversations despite each turn providing relevant material. If $\gamma = 1$ then the metric ignores the dependency between turns – and can be considered a binary version of CCG.

To provide some examples of how the metric is calculated, imagine we have the following conversational paths where we observe path A, $p(A) = \{r(t_1, t_2), n(t_3), r(t_4), n(t_5)\}$ and path B, $p(A) = \{n(t_1), r(t_2, t_2, t_3), n(t_5)\}$. Both paths have three turns which are considered reasonably relevant responses. However, path B has 3 relevant turns in a row, while path A has 2 relevant turns in a row, and then an isolated relevant turn. The CPS for A then is $(2^\gamma + 1^\gamma)/5^\gamma)$, while for B it would be $(3^\gamma)/5^\gamma$. If $\gamma = 2$ then A would receive a score of 5/25 while B would receive a score of 9/25. And, so with a $\gamma > 1$ we can see that path B receives the higher score.

Our next metric, **Turn-Biased Conversational Cumulative Gain** (TBCCG) provides a different variation on the former two approaches by encoding an explicit user "conversing" model[3] into the metric. After each turn, we assume that a user will either continue the conversation or stop. The probability of the user continuing depends upon whether the system's response was sufficiently relevant. We further assume that the probability of continuing to give a relevant response is greater than the probability of continuing to give a non-relevant response. And as above, if the score for the turn is above the threshold θ, i.e., $s_t > \theta$ then we assume the response was reasonably relevant r, else non-relevant n. Essentially, this means that only a certain proportion of users will reach subsequent turns – depending on the probabilities of continuing and the relevance of the responses.

To compute TBCCG, we first need a vector of continuation probabilities associated with each subsequent turn to calculate the metric. In the first turn, we assume all users consider the response and then depending on the response, the user continues with some probability $P(c|t_1)$. The proportion of users that consider the first turn is 1.0, the proportion of users that consider the second turn is $1.0 \times P(c|t_1)$, the third turn $1.0 \times P(c|t_1) \times P(c|t_2)$, and so on. For a given path p where we have turns t_1 to t_n, we can calculate the TBCCG as follows:

$$TBCCG(s) = \frac{1}{|p|}\left(s_{t_1} + \sum_{i=2}^{n} \prod_{i=2}^{n} P(c|t_{i-1}).s_{t_i}\right) \tag{3}$$

where s_{t_i} is the score for the ith turn, and the probability of continuing $P(c|t_i)$ depends on the relevance of the response, where $P(c|t_i) = P_r$ if $s_{t_i} > \theta$, else $P(c|t_i) = P_n$. We assume that $P_r \geq P_n$ is such that users are more likely to continue the conversation when they receive a relevant response than when they receive a non-relevant response. Note that if the probability of continuing is for $P_r = P_n = 1$ then the metric reverts back to CCG. If p_n is set to 0, then, as soon as a user encounters a non-relevant turn, then they would stop, and so the user would not accumulate any further gain. While if P_n is greater than zero, it suggests there is some probability that the user will persist and continue interacting.

$$WTBCCG(s) = \frac{1}{\sum_i w_{t_i}}\left(w_{t_1}.s_{t_1} + \sum_{i=2}^{n} \prod_{i=2}^{n} P(c|t_{i-1}).w_{t_i}.s_{t_i}\right) \tag{4}$$

Response Quality Evaluation: On top of the standard provenance ranking, we also evaluate the response quality of the main task runs in terms of their relevance, naturalness, and conciseness. Judgments for these dimensions are collected as described in 2.1. We use Precision @ 1 and NDCG @ 1 as a proxy for these evaluations.

3.2 MI Sub-task

We evaluate the MI submissions in two ways, namely, human evaluation and end-to-end evaluation.

Human evaluation. As described in Section 2.2, we collect human judgments on the MI submissions, concerning different aspects of the utterance, namely, relevance, novelty, and diversity. We evaluate the runs in terms of collected relevance labels and report the results in terms of NDCG@1 and P@1. We take only

[3] Akin to the user browsing model when interacting with a ranked list.

Table 2: Participants and their runs.

Group	Run ID	Run Type	Group	Run ID	Run Type
CFDA_CLIP	CNC_AD	automatic	MLIA-DAC	MLIA_DAC_splade	automatic
CFDA_CLIP	CNC_AD-C	automatic	MLIA-DAC	splade_t5mm	automatic
CFDA_CLIP	CNC_AS	automatic	MLIA-DAC	splade_t5mm_ens	automatic
CFDA_CLIP	CNC_AS-C	automatic	MLIA-DAC	splade_t5mse	automatic
CFDA_CLIP	CNC_MD-C	manual	udel_fang	udinfo_best2021	automatic
CFDA_CLIP	CNC_MS-C	manual	udel_fang	udinfo_mi_b2021	automatic-mi
CNR	CNR_run1	automatic	udel_fang	udinfo_onlyd	automatic
CNR	CNR_run2	manual	udel_fang	udinfo_onlyd_mi	automatic-mi
CNR	CNR_run3	automatic	UiS	uis_cargoboat	automatic
CNR	CNR_run4	automatic	UiS	uis_duoboat	automatic
HEATWAVE	combine0.5	automatic	UiS	uis_mixedboat	automatic-mi
HEATWAVE	duo_reranker	automatic	UiS	uis_sparseboat	automatic
HEATWAVE	gold	manual	uogTr-AT	UoGTr	automatic
HEATWAVE	monot5	automatic	uogTr-MI-HB	UoGTr	automatic-mi
iiia-unipd	DEI-run1	automatic	uogTr-MT	UoGTr	manual
iiia-unipd	DEI-run2	automatic	WaterlooClarke	UWCauto22	automatic
iiia-unipd	DEI-run4	automatic	WaterlooClarke	UWCcano22	automatic
iiia-unipd	DEI-run5.json	automatic	WaterlooClarke	UWCmanual22	manual

the top one utterance, as we assume that only one MI utterance is posed to the user.

End-to-end evaluation. We give teams who participate in the MI sub-task the opportunity to leverage the MI turns in the main task, i.e., passage retrieval. Therefore, we also evaluate the MI submissions in an end-to-end manner where we examine how much it affects the final passage ranking performance. Note that this is not an accurate measure of the quality of MI submissions, since the quality of the ranker (and how it incorporates MI turns) plays a role in the performance.

4 PARTICIPANTS

4.1 Main Task

The CAsT main task received 36 run submissions from 9 groups shown in Table 2. The organizers provided two runs (one automatic, one manual) as baselines for comparison. Participants provided metadata and descriptions of their runs.

Similar to previous years, many teams used a multi-step pipeline consisting of: (1) conversational rewriting (most incorporating the previous canonical responses) and conversational query expansion, (2) retrieval using traditional IR or dense model, and (3) multi-stage passage re-ranking with neural language models fine-tuned for point-wise (mono) and pairwise (duo) ranking. Almost all teams leverage pre-trained Transformer-based language models for rewriting (BART, T5) and ranking (mostly T5). There continued to be a trend of using learned representations for ranking, including sparse (Splade) and dense retrieval. There's also an emerging thread on dense conversational query expansion. In practice, many of the best-performing runs appeared to use a fusion of retrieval approaches for first-pass retrieval and combined re-ranking scores from the multiple passes of retrieval.

4.2 MI Sub-task

The sub-task received 10 runs across 5 groups. In addition, the organizers provided four additional baseline runs (described in 2.1). We saw a variety of approaches used for the task. All runs used some form of ranking, generative, and template-based approaches, with some using a combination of the three. Generative approaches mostly used the ClariQ dataset for training and almost all of these runs use a T5 model for clarification generation. The UoG_GRILL used GPT-3 with few-shot prompting from 2021 CAsT data. Most runs asked questions for all turns, and only four of the runs performed selective MI generation.

5 OVERALL RESULTS

In this section, we present the results of the submitted runs. We include the organizer baselines (*Organizers* Group) described above that are available in the public CAsT Github repository[4].

5.1 Main Task

For the main task, we evaluate the ranking of the provenance passages. The results are turn-level macro-averaged retrieval effectiveness. We use four standard evaluation measures: Recall, Mean Average Precision (MAP@1K), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@1K). The primary measure continues to be NDCG@3, focusing on high-precision and quality responses in the top ranks. We use a relevance cutoff of **two** as positive for binary measures because the value of one is marginal accordingly to the guidelines.

We distinguish between the two broad categories of runs: *automatic* and *manual*. Automatic runs use the raw utterances or provided automatic rewrites. The Automatic MI runs also use responses from the MI sub-task. Manual runs use manually rewritten (resolved) queries.

[4] https://github.com/daltonj/treccastweb

Table 3: Automatic evaluation of provenance retrieval results. Evaluation at retrieval cutoff of 1000 with a binary relevance threshold of 2.

Group	Run	Recall	MAP	MRR	NDCG	NDCG@3
udel_fang	udinfo_mi_b2021	**0.771**	**0.246**	**0.656**	**0.557**	**0.452**
udel_fang	udinfo_onlyd_mi	0.729	0.243	0.646	0.540	0.450
HEATWAVE	duo_reranker	0.453	0.189	0.639	0.392	0.440
HEATWAVE	combine0.5	0.453	0.184	0.641	0.389	0.439
HEATWAVE	monot5	0.453	0.178	0.619	0.381	0.426
WaterlooClarke	UWCcano22	0.556	0.215	0.624	0.445	0.424
MLIA-DAC	splade_t5mm_ens	0.638	0.219	0.592	0.483	0.416
MLIA-DAC	splade_t5mm	0.638	0.202	0.574	0.470	0.401
UiS	uis_sparseboat	0.507	0.189	0.566	0.409	0.388
CFDA_CLIP	CNC_AS	0.527	0.184	0.559	0.411	0.377
WaterlooClarke	UWCauto22	0.517	0.206	0.566	0.416	0.377
UiS	uis_cargoboat	0.450	0.180	0.526	0.377	0.373
UiS	uis_mixedboat	0.445	0.186	0.499	0.374	0.363
—	BM25_T5_BART_automatic	0.324	0.150	0.527	0.299	0.362
MLIA-DAC	MLIA_DAC_splade	0.638	0.162	0.514	0.433	0.348
udel_fang	udinfo_onlyd	0.651	0.178	0.531	0.453	0.348
CFDA_CLIP	CNC_AD	0.320	0.117	0.517	0.294	0.347
UiS	uis_duoboat	0.365	0.154	0.476	0.323	0.345
CFDA_CLIP	CNC_AD-C	0.320	0.109	0.487	0.286	0.334
UoGTr	uogTr-MI-HB	0.413	0.152	0.503	0.337	0.331
udel_fang	udinfo_best2021	0.681	0.181	0.514	0.465	0.325
UoGTr	uogTr-AT	0.319	0.134	0.494	0.290	0.317
iiia-unipd	DEI-run1	0.445	0.126	0.443	0.327	0.280
iiia-unipd	DEI-run5	0.420	0.120	0.455	0.315	0.276
iiia-unipd	DEI-run4	0.456	0.126	0.441	0.334	0.274
MLIA-DAC	splade_t5mse	0.638	0.127	0.410	0.393	0.271
iiia-unipd	DEI-run2	0.458	0.125	0.426	0.333	0.263
CFDA_CLIP	CNC_AS-C	0.515	0.141	0.366	0.369	0.235
CNR	CNR_run4	0.316	0.072	0.367	0.224	0.216
CNR	CNR_run3	0.332	0.068	0.340	0.231	0.201
CNR	CNR_run1	0.333	0.059	0.274	0.219	0.177

Table 4: Manual provenance ranking results. These runs used the manually resolved queries. Evaluation at retrieval cutoff of 1000 with a binary relevance threshold of 2.

Group	Run	Recall	MAP	MRR	NDCG	NDCG@3
HEATWAVE	gold	0.557	0.257	**0.717**	0.485	**0.513**
CFDA_CLIP	CNC_MD-C	0.339	0.163	0.706	0.350	0.512
—	BM25_T5_BART_manual	0.465	0.231	0.716	0.423	0.503
WaterlooClarke	UWCmanual22	0.676	0.294	0.711	**0.550**	0.501
UoGTr	uogTr-MT	0.553	0.249	0.695	0.477	0.487
CFDA_CLIP	CNC_MS-C	**0.702**	**0.260**	0.593	0.537	0.398
CNR	CNR_run2	0.382	0.076	0.338	0.255	0.202

Automatic run results. Table 3 shows the results for the 31 automatic runs with a median NDCG@3 score of 0.348.

The best-performing run uses fusion on top of four retrieval methods that each include sparse and dense retrieval combined with mono-duo T5 reranking. It also includes sparse and dense conversational query expansion. It also incorporates output from MI interactions. The second best-performing run only uses dense retrieval.

Manual run results. Table 4 shows the results for the seven manual runs with a median NDCG@3 value of 0.501. Two runs outperform the organizer bm25_t5 re-ranking baseline. One of these, *CNC_MD-C*, uses conversational dense passage retrieval (ConvDPR) and mono reranking. The highest recall is 0.702 and uses sparse retrieval. The UWCmanual22 also achieves high recall and is based on feedback from external corpora.

Overall. It is noteworthy that two automatic runs achieve high recall (> 0.7) and are more effective in recall than any of the manual runs. However, the automatic results still lag behind in terms of precision in the top ranks. Although the gap in candidate retrieval appears to have shrunk (or even closed), using manual queries in reranking still results in significant gains (over 13% relative gain in the best runs) over the best automatic method. We also observe that the best run uses mixed-initiative interaction data.

Results by Topic: Figure 1 provides a per-topic analysis comparing the three classes of systems across topics. It uses data from runs above the median. The results show that the topic difficulty varies widely across topics. Interestingly, automatic-mi runs perform better than manual runs 4 / 17 topics (25%). There are 11 topics where manual runs still strongly outperform automatic and automatic-mi methods (over half). Topics 133, 142, and 143 have the largest gap between manual and automatic runs. The hardest topics across all types are 135 and 144 which include several feedback and comment turns.

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 eleven are truncated. Figure 2 shows the average NDCG@3 at each turn depth for the categories of runs. The figure only shows the data for runs that perform at or above the median NDCG@3.

5.1.1 Conversational Path Evaluation.

In this section, we go beyond turn-level metrics and focus on evaluating the conversations holistically across turns. Given our three proposed related conversational metrics, we instantiated a number of different variations, but only report a small subset. For those reported, we use the official turn level metric NDCG@3 as s given the output of the strict evaluation (relevance of 2 or higher is relevant). We then set the threshold θ of 0.33 for all reported metrics. That is we assume a reasonably relevant response would return at least one highly relevant item, one relevant and one marginally relevant item, or three marginally relevant items. For our CPS metric, we set γ to 2 and 3, and for TBCCG we fixed P_r equal to one (always continue if relevant) and then varied P_n was set to 0.0 and 0.25. The mean of each metric is reported. Note that this is the mean over all the conversational tree paths, not the mean over turns.

Table 5 presents the results given our conversational path measures where systems have been ordered by CCG. It is worth noting that the correlation between our conversational path measures is very high (with Pearson's r of approx. 0.88-0.97). As a result, the systems with high CCG also obtained high scores on our conversational flow-based measures. Over the conversational path, measures present a slightly different story than turn-based ones. The top ranking system according to the CCG, CPS, and TBCCG is the HEATWAVE run combine0.5. However, when the flow is weighted more highly (i.e. CPS with $\gamma = 3$) then we see that udel_-fang's run udinfo_mi_b2021 is more effective. Clearly, the more turns that are considered as reasonably relevant responses ($s_t \geq \theta$), the higher their conversational path metric scores are likely to be — and the higher the penalty to runs which can not consistently return longer sequences of relevant responses. In contrast with the pure turn-based evaluation, we see a difference in the order of the top runs – suggesting that if we also care about consistency of and flow of experience, simply maximizing turn-based NDCG@3 on its own is not optimal.

Table 6 presents the results for the manual runs for our conversational path measures. Here, we see that CFDA CLIP's CNC MD-C run consistently outperforms all the other automatic runs – which is also consistent with the turn-based NDCG@3 results. It changes the order of the second and third teams, with UWCmanual22 outperforming the organizer baseline for all the measures. We also observe that the manual runs above the median outperform all automatic runs on CCG. It also shows that even for manual runs there is still significant headroom to focus on consistency.

5.1.2 Response Quality.

In this section, we describe the response quality output using crowdsourced judgments. Given that all responses have at least one judged relevant document in their provenance, the focus is on how this is presented to the user in the response. The results tables are sorted by the mean of all three factors.

This evaluation is performed only on the subset of queries where the system returned a response with a relevant document in its top three responses. This focuses the evaluation on the quality of the response, not on core relevance, which is already reflected in the main task turn relevance evaluation. Note that the UoGTr group runs are not included in the response quality evaluation because they were added after the quality assessment was performed.

Table 7 shows the quality of the automatic results. We observe that the perceived response relevance is high overall, as we would expect given that at least one relevant passage is being used in the response, but there is wide variance. The best-performing conciseness is from generative models, the WaterlooClarke runs used T5 and the Organizer run used BART to generate abstractive summaries of the results. Examining the results, we observe that QA models produced very short segments, that did not score as well for relevance and naturalness. Systems that returned long passages (up to 250 words) were sometimes judged to be less natural as well as less concise. Table 8 shows the same response quality measures for the manual runs. The relevance is comparable to the automatic runs as well as conciseness and naturalness.

Overall, the effectiveness for manual and automatic runs are both low around turns four and eight with the high points at five and nine. Over time we see a decline in the effectiveness of automatic systems. We also studied the differences in automatic and automatic-mi runs by depth and we observe that turns 5 and 7 are higher for automatic-MI runs and comparable for the other turns.

The gap between manual and automatic is smaller than in previous years, with a smaller gap between them even later in the conversation. The number of turns at depths 10 and 11 is quite limited and not enough to draw conclusions.

5.2 MI Sub-Task

In this section, we present the results of the MI sub-task both in terms of human evaluation.

Human Intrinsic Relevance Evaluation There are human annotations that include relevance as well as other MI factors. In this

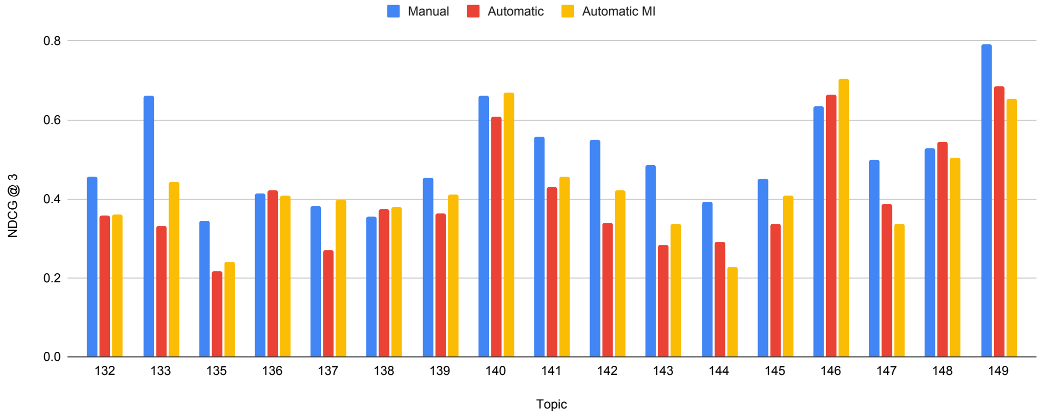

Figure 1: NDCG@3 aggregated for each topic across all runs. We report the average across runs, median or better for each run category.

Table 5: Conversational Path Evaluation on the Main Task for Automatic Runs. The turn-based metric used was NDCG@3 with a threshold $\theta = 0.33$.

	run	CCP	CPS($\gamma = 2$)	CPS($\gamma = 3$)	TBCCG($P_n = 0.0$)	TBCCG($P_n = 0.25$)
HEATWAVE	combine0.5	**0.341**	**0.224**	0.131	**0.138**	**0.171**
udel_fang	udinfo_mi_b2021	0.334	0.215	**0.139**	0.110	0.144
HEATWAVE	duo_reranker	0.327	0.185	0.102	0.122	0.151
udel_fang	udinfo_onlyd_mi	0.326	0.175	0.097	0.091	0.128
MLIA-DAC	splade_t5mm_ens	0.313	0.139	0.060	0.083	0.114
WaterlooClarke	UWCcano22	0.307	0.181	0.107	0.093	0.124
MLIA-DAC	splade_t5mm	0.294	0.126	0.054	0.071	0.103
UiS	uis_sparseboat	0.291	0.155	0.077	0.086	0.118
CFDA_CLIP	CNC_AS	0.290	0.167	0.091	0.112	0.140
UiS	uis_cargoboat	0.281	0.116	0.046	0.090	0.119
WaterlooClarke	UWCauto22	0.275	0.121	0.056	0.079	0.110
-	BM25_T5_BART_automatic	0.273	0.134	0.061	0.093	0.120
MLIA-DAC	MLIA_DAC_splade	0.270	0.099	0.033	0.053	0.082
CFDA_CLIP	CNC_AD-C	0.269	0.107	0.044	0.093	0.116
HEATWAVE	monot5	0.267	0.141	0.067	0.092	0.119
CFDA_CLIP	CNC_AD	0.266	0.146	0.066	0.110	0.133
UiS	uis_mixedboat	0.264	0.100	0.038	0.072	0.096
udel_fang	udinfo_onlyd	0.258	0.116	0.061	0.076	0.099
UiS	uis_duoboat	0.257	0.104	0.040	0.067	0.096
udel_fang	udinfo_best2021	0.253	0.110	0.057	0.080	0.102
UoGTr	uogTr-MI-HB	0.247	0.101	0.039	0.077	0.098
UoGTr	uogTr-AT	0.245	0.113	0.050	0.074	0.099
MLIA-DAC	splade_t5mse	0.221	0.081	0.028	0.079	0.095
iiia-unipd	DEI-run1	0.218	0.065	0.019	0.049	0.070
iiia-unipd	DEI-run4	0.212	0.065	0.019	0.051	0.070
iiia-unipd	DEI-run2	0.210	0.058	0.017	0.049	0.068
iiia-unipd	DEI-run5	0.205	0.061	0.019	0.060	0.077
CFDA_CLIP	CNC_AS-C	0.189	0.080	0.034	0.053	0.071
CNR	CNR_run4	0.175	0.076	0.037	0.052	0.067
CNR	CNR_run3	0.160	0.054	0.018	0.035	0.051
CNR	CNR_run1	0.144	0.057	0.021	0.030	0.047

Table 6: Conversational Path Evaluation on Main Task over Manual Runs. These runs used manually resolved queries. The turn-based metric used was NDCG@3 with a threshold $\theta = 0.33$.

	run	CCG	CPS($\gamma = 2$)	CPS($\gamma = 3$)	TBCCG($P_n = 0.0$)	TBCCG($P_n = 0.25$)
CFDA_CLIP	CNC_MD-C	**0.389**	**0.264**	**0.169**	**0.154**	**0.193**
WaterlooClarke	UWCmanual22	0.370	0.222	0.122	0.131	0.168
-	BM25_T5_BART_manual	0.369	0.205	0.096	0.126	0.164
UoGTr	uogTr-MT	0.358	0.198	0.114	0.140	0.166
HEATWAVE	gold	0.325	0.186	0.095	0.105	0.140
CFDA_CLIP	CNC_MS-C	0.294	0.180	0.117	0.073	0.103
CNR	CNR_run2	0.159	0.054	0.017	0.032	0.047

Table 7: Response quality evaluation for automatic main task runs using crowdsourced jugdments.

	Run	Relevance @ 1	Conciseness @ 1	Naturalness @ 1
WaterlooClarke	UWCauto22	0.803	**0.667**	**0.704**
WaterlooClarke	UWCcano22	0.807	0.591	0.591
Organizers	BM25_T5_BART_automatic	0.745	0.617	0.500
udel_fang	udinfo_mi_b2021	0.820	0.470	0.530
udel_fang	udinfo_best2021	0.779	0.456	0.529
MLIA-DAC	splade_t5mse	**0.842**	0.439	0.456
UiS	uis_duoboat	0.809	0.412	0.515
udel_fang	udinfo_onlyd_mi	0.796	0.439	0.490
UiS	uis_sparseboat	0.803	0.382	0.434
UiS	uis_cargoboat	0.836	0.315	0.466
udel_fang	udinfo_onlyd	0.716	0.419	0.446
UiS	uis_mixedboat	0.761	0.375	0.421
HEATWAVE	combine0.5	0.634	0.430	0.484
CFDA_CLIP	CNC_AD	0.645	0.355	0.419
CFDA_CLIP	CNC_AD-C	0.568	0.432	0.398
HEATWAVE	duo_reranker	0.557	0.454	0.361
CFDA_CLIP	CNC_AS	0.621	0.359	0.388
MLIA-DAC	splade_t5mm	0.700	0.338	0.325
HEATWAVE	monot5	0.539	0.393	0.427
MLIA-DAC	splade_t5mm_ens	0.713	0.310	0.322
CNR	CNR_run3	0.695	0.237	0.390
MLIA-DAC	MLIA_DAC_splade	0.722	0.250	0.306
CNR	CNR_run4	0.657	0.271	0.343
CNR	CNR_run1	0.729	0.188	0.354
iiia-unipd	DEI-run4	0.491	0.309	0.400
iiia-unipd	DEI-run5	0.462	0.269	0.365
CFDA_CLIP	CNC_AS-C	0.480	0.370	0.233
iiia-unipd	DEI-run2	0.481	0.250	0.346
iiia-unipd	DEI-run1	0.482	0.259	0.333

Table 8: Response quality evaluation for manual main task runs using crowdsourced jugdments.

	Run	Relevance @ 1	Conciseness @ 1	Naturalness @ 1
WaterlooClarke	UWCmanual22	**0.856**	**0.673**	**0.673**
Organizers	BM25_T5_BART_manual	0.744	0.566	0.566
CFDA_CLIP	CNC_MD-C	0.623	0.485	0.492
HEATWAVE	gold	0.642	0.406	0.453
CFDA_CLIP	CNC_MS-C	0.505	0.367	0.358
CNR	CNR_run2	0.688	0.203	0.266

Table 9: Clarification question evaluation of MI runs using crowdsourced jugdments.

	Run	Relevance @ 1	Novelty @ 1	Diversity @ 1
UoG_GRILL	GPT-3_full_context	**0.852**	**0.536**	**0.607**
UoG_GRILL	GPT-3_rewrite	0.657	0.515	0.551
UoGTr	uogTr-MI	0.663	0.494	0.369
UiS	uis_clearboat	0.639	0.488	0.371
UoG_GRILL	GPT-3_raw	0.592	0.362	0.490
Organizers	miniLM_bert_sample_mi_run	0.371	0.317	0.395
Organizers	bm25_baseline_mi	0.345	0.293	0.307
UiS	uis_vagueboat	0.237	0.322	0.381
CFDA_CLIP	CNC_kwqlm2_cqg	0.340	0.283	0.220
CFDA_CLIP	CNC_kwqlm_cqg	0.340	0.263	0.234
Organizers	T5_rewrite	0.320	0.229	0.210
CFDA_CLIP	CNC_cqg	0.294	0.234	0.224
UDel	mi_task_0822_1	0.155	0.117	0.322
Organizers	T5_raw	0.232	0.166	0.185

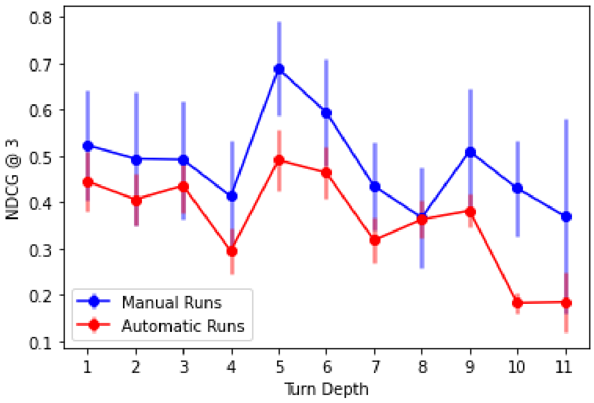

Figure 2: NDCG@3 at varying conversation turn depths. We report the average across runs, median or better.

section, we focus just on the relevance of the posed question to the previous response using crowdsourced judgments. Table **??** shows the human evaluation of the MI submissions. The table includes all turns as well as predicted turns, a subset where the system

We evaluate the submissions in two settings, namely, on all the turns and only the utterances where the system posed MI questions. When evaluating on all the turns, we see that the *uis-clearboat* achieves the highest NDCG@1 and P@1. It's significantly higher than the other approaches. However, when we focus only on the subset of turns where MI questions are posed, the GPT-3 method by UofG_GRILL is the most effective, with an almost 20% absolute difference in NDCG@1 over the next best system. This approach implicitly performs classification of when to use MI by including it in the few-shot prompt. The uogTR-MI run also performs well in both and incorporates a retrieve and generate approach along with further T5 question ranking.

6 CONCLUSION

The fourth TREC CAsT edition developed resources for studying conversational information seeking and added to the community's understanding of the topic. It made significant advances over the previous edition to focus on generating responses, having multiple varied conversations on the same topic using topic trees, and becoming more realistic and interactive.

The MI sub-task provided a way for participants to gain interactive feedback to improve effectiveness. The most effective automatic team leveraged MI responses. However, the static nature of the topics for the main task limited the headroom of these methods. Future directions should incorporate these into topic development more deeply and support multiple rounds of MI as the conversations evolve.

This year also introduced grounded response generation to focus not just on retrieving content, but on synthesizing relevant, concise, and natural responses. This is an important emerging area that will continue to grow in importance as rapid advancements in generative language models become more widely used in conversational information-seeking systems.

We look forward to future interactive tracks that continue to push the boundary of systems that can meaningfully interact in order to help people accomplish increasingly complex tasks.

7 ACKNOWLEDGMENTS

We thank Carlos Gemmell, Iain Mackie, and Alessandro Speggorin for their contributions to topic development and annotation. We also are deeply thankful for Ian Soboroff's experience, patience, and persistence in running the assessment process. Finally, we thank all our participants. The crowd-sourcing annotation efforts are supported by EPSRC, the grant, Neural Conversational Information Seeking Assistant, EP/V025708/1.

REFERENCES

[1] Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.S.: Building and evaluating open-domain dialogue corpora with clarifying questions. In: EMNLP (1). pp. 4473–4484. Association for Computational Linguistics (2021)

[2] Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: SIGIR. pp. 475–484. ACM (2019)

[3] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Lin, J.: Overview of the trec 2021 deep learning track. In: Text REtrieval Conference (TREC). TREC (May 2022), https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/

[4] Dalton, J., Xiong, C., Callan, J.: TREC CAsT 2021: The conversational assistance track overview. In: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021). National Institute of Standards and Technology, special publication (2021)

[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)