University of Glasgow Terrier Team at the TREC 2022 Conversational Assistance Track

Sarawoot Kongyoung
University of Glasgow, UK
s.kongyoung.1@research.gla.ac.uk

Craig Macdonald
University of Glasgow, UK
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow, UK
iadh.ounis@glasgow.ac.uk

ABSTRACT

In this paper, we describe our methods and submitted runs for the TREC 2022 Conversational Assistance Track. In our participation, we leverage Multi-Task Learning (MTL) methods to enhance the performance of the conversational search system. For the main task, we use our recently proposed monoQA model, which applies Multi-Task Learning (MTL) on reranking and answer extraction by sharing a single text generation model, predicts both the answer and the reranking score simultaneously. For the mixed-initiative sub-task, we propose T5MI, which is trained on the ClariQ dataset, to determine whether a user utterance needs to ask clarifying questions, as well as to generate useful clarifying questions. This year, we submitted three runs based on the data used in the testing step consisting of 1) uogTr-MT: using the provided manually rewritten utterances as the queries; 2) uogTr-AT: using the raw utterances and the provided provenances as the context for rewriting the queries; 3) uogTr-MI-HB: using the raw utterances and the output from the mixed-initiative sub-task as the context for rewriting the queries.

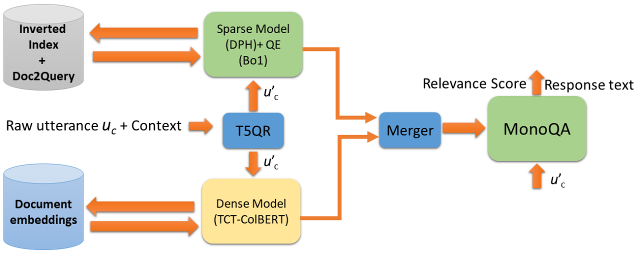
Inverted Index + Doc2Query
Sparse Model (DPH)+ QE (Bo1)
u'_c
Raw utterance u_c + Context
T5QR
u'_c
Document embeddings
Dense Model (TCT-ColBERT)
Merger
MonoQA
Relevance Score
Response text
u'_c

Figure 1: The overall architecture of our system.

1 INTRODUCTION

CAsT 2022 is the fourth year of the Conversational Assistance Track in TREC. The CAsT track tackles information retrieval tasks in a conversational context. Similar to previous years, the canonical responses to each user's utterance and explicit feedback are provided. The main difference with last year's setup is that a new sub-task focusing on mixed-initiative has been added. For each turn in the conversation, the system may give a response or ask a question.

In this work, to address the conversational search task of the Conversational Assistance track, we followed a multi-stage framework consisting of a query rewriting, a query and document expansions, a retriever, a reranker, and a reader, as illustrated in Figure 1. In particular, we leverage Multi-Task Learning (MTL) consisting of monoQA (our recently proposed MTL model, which combines reranking and answer extraction) [6] and T5MI (which applies MTL of clarification need classification and clarifying question generation) to address the conversational search task. Firstly, the query rewriting model takes the raw user utterance and its context as an input sequence and reformulates it into a fully specified query. Secondly, each document is expanded before indexing and each query is expanded before feeding it into the retriever. Thirdly, the retriever retrieves the top K relevant passages from the text collection based on a query rewritten by the query rewriting model. The reranker and the reader then respectively rerank and identify an answer in the top K passages.

The structure of the remainder of this paper is structured as follows: Section 2 discusses our multi-stage pipeline setup; Section 3 describes our mixed-initiative method; Section 4 describes our submitted runs; Section 5 reports our results; Concluding remarks follow in Section 6.

2 A MULTI-STAGE PIPELINE FOR A CONVERSATIONAL SEARCH SYSTEM

In this section, we describe our multi-stage pipeline to address the conversational search task as illustrated in Figure 1. In the following, we describe each of these stages in detail.

2.1 T5 Query Rewriting Model

To deal with the ambiguity of conversational questions, we use a T5 [17] query rewriting (T5QR) model, which has been fine-tuned using the CANARD [3] conversational question rewriting dataset. By doing this, we follow the CAsT baseline rewriting configuration by using all historical utterances $u_{1:k-1}$ and all canonical response passages $r_{1:k-1}$ as the context. For example, the specific user utterance at turn k (u_k) can be reformulated as follows:

$$u_k^{'} = T5QR(u_1|||r_1|||...|||u_{k-1}|||r_{k-1}|||u_k) \tag{1}$$

where $u_k^{'}$ is the rewritten query, which corresponds to the raw utterance u_k and its context ($u_{1:k-1}; r_{1:k-1}$). '|||' is used to separate tokens.

2.2 Query & Document Expansions

Query expansion: Following our participation in the TREC 2021 Deep Learning track [19], we use the Bo1 [2] query expansion method to improve the rewritten queries for document retrieval.
Document expansion: Following doc2query [13], we employ document expansion with predicted queries. In particular, for each document in the three collections (MS MARCO V2, KILT, WaPo), we add three predicted queries to the text of the document:

$$d^{'} = d \oplus (\hat{q}_1 \oplus \hat{q}_2 \oplus \hat{q}_3) \tag{2}$$

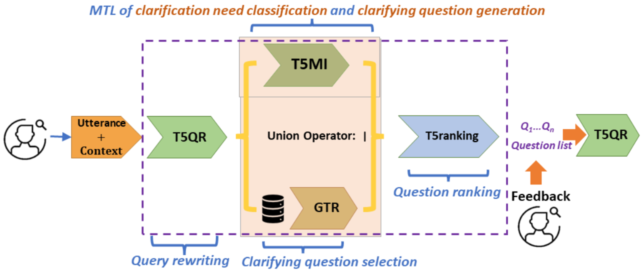

Figure 2: The overall framework of our mixed-initiative system.

where d and $d^{'}$ are, respectively, the original document and the expanded document that is obtained by appending the predicted queries $\hat{q}_{1:3}$. We apply the docTTTTTquery [13] model to generate a predicted query from each document.

2.3 Hybird of Sparse and Dense Retrieval

For passage retrieval, following our participation in the TREC 2021 Deep Learning track [19], we implement a hybrid of sparse and dense retrieval model as shown in Figure 1. In the hybrid of sparse and dense retrieval pipeline, we combine spare retrieval, namely DPH with the Bo1 [2] query expansion mechanism on the inverted index with TCT-ColBERT [8, 10] on the FAISS [4] index. In order to create a hybrid of the sparse and dense retrieval models, we take the union of the passages returned by the dense retrieval model and those returned passages by the sparse retrieval model. The merged passages are then reranked and the answer is extracted using the monoQA model, which is described next.

2.4 monoQA: Reranker & Generative Reader

In our participation this year, we deploy our recently proposed monoQA [6] MTL model, which is fine-tuned simultaneously for both reranking (in order to improve the precision of the top retrieved passages) and extracting the answer on the OR-QuAC dataset [16]. Our MTL-based monoQA model has been shown to outperform several existing multi-stage pipeline systems, such as the ORConvQA systems proposed by [7, 9, 15, 16] or the separate applications of the monoT5 [12] and UnifiedQA [5] models. In particular, compared to the separate applications of the monoT5 and UnifiedQA models for reranking and extracting the answer, our MTL-based monoQA model is twice as fast for inference. This motivates our use of monoQA in our participation in the CAsT Track this year.

3 MIXED-INITIATIVE SUB-TASK

In this section, we describe in detail our method and system implementation to address the mixed-initiative sub-task.

3.1 T5MI: Clarification Need Classification & Clarifying Question Generation

The upper part of Figure 2 shows our proposed model, which uses the T5 model, a large pre-trained language model designed for text generation, namely T5MI. We leverage Multi-Task Learning with a text generation model to effectively address the tasks of clarification need classification and clarifying question generation. Our T5MI model aims to identify whether a user utterance requires a clarifying question and, accordingly, generates a clarifying question using the rewritten question from the query rewriting model.

Table 1: Pipeline components used for each run

Run	QR	Retrieval	QE & DE	Re-ranker & Reader
		official runs		
uogTr-MT	provided manually rewritten utterances	DPH & TCT-ColBERT	Bo1 + doc2query	monoQA
uogTr-AT	raw utterances+ full context	DPH & TCT-ColBERT	Bo1 + doc2query	monoQA
uogTr-MI-HB	raw utterances+ output from the sub-task	DPH & TCT-ColBERT	Bo1+ doc2query	monoQA

3.2 Question Selection

As shown in the bottom part of Figure 2, we use a Generalizable T5-based dense Retriever (GTR) model [11] to retrieve the clarifying questions from the questions pool provided by the organisers. By doing this, we use the provided checkpoint `gtr-t5-xxl`, without further fine-tuning.

3.3 Question Ranking

To rank the clarifying questions obtained from both the T5MI and the GTR models, we score them using a T5 model, denoted as T5Ranking, trained on the ClariQ [1] dataset for pointwise question classification, as illustrated in the right part of Figure 2. Once the ranking of questions was submitted to TREC, for each turn, the organisers collected user feedback for the top-one question from the question ranking. We then use the provided feedback for each clarifying question as the context for rewriting the raw utterance.

4 RUNS

In this section, we describe our submitted run for the Mixed-Initiative sub-task in Section 4.1. The details of our submitted runs for the Conversational Search task are provided in Section 4.2.

4.1 Mixed-Initiative Sub-task

Official Runs

- **uogTr-MI:** Applies the T5 query rewriting (T5QR) model using the full context of the provided canonical responses and all historical utterances to rewrite the current utterance. Next, the rewritten utterance is used as input for the MTL T5MI model to identify whether the utterance needs to ask a clarifying question, as well as to generate clarifying questions. In addition, to retrieve the clarifying question from the provided question pool, we adopt a GTR dense retrieval for indexing and retrieving the relevant questions. Finally, we combine the generated and retrieved clarifying questions and rank them using the T5Ranking model.

Provided Baselines

For comparison with our results in the mixed-initiative sub-task, we include the results of the baseline methods provided by the track organisers [14], namely:

Table 2: Results on the TREC Conversational Assistance Track 2022 Mixed-Initiative sub-task. The best performing run for each measure is emphasised.

Approach	Relevance@1	Novelty@1	Diversity@1
generation baselines			
GPT-3 Raw	0.433	0.263	0.356
GPT-3 Rewrite	0.454	0.346	0.371
GPT-3 Full	0.119	0.073	0.082
T5 Raw	0.232	0.166	0.185
T5 Rewrite	0.320	0.229	0.210
selection baselines			
BM25	0.345	0.293	0.307
miniLM-BERT	0.371	0.317	**0.395**
hybrid of generation and selection			
uogTr-MI	**0.567**	**0.494**	0.369

Generation: The generation baseline runs use GPT3 and T5 with different types of input to generate questions. The T5 model was finetuned on the ClariQ dataset:

- GPT-3 Raw: Uses the raw user utterance at each turn as input to GPT3.
- GPT-3 Rewrite: Uses the automatic rewrite as input to GPT3.
- GPT-3 Full: Uses the full conversation history (user utterances and system responses) as input to GPT-3.
- T5 Raw: Uses the raw user utterance as input to T5.
- T5 Rewrite: Uses the automatic rewrite as input to T5.

Selection: The selection baseline runs rank questions from the question pool:

- BM25: Ranks questions using BM25 with the automatic rewrite of the current turn query.
- miniLM-BERT: Generates a candidate pool of questions using the `all-MiniLM-L6-v2` model [18] from Sentence Transformers, then reranks them using a BERT model trained on the ClariQ dataset for pointwise question classification.

4.2 Conversational Search Task

We submitted 3 runs to the Conversation Search task. Table 1 describes the features we used in each of our runs.
Official Runs

- **uogTr-MT:** Produces a run using the manually rewritten utterances provided by the track organisers.
- **uogTr-AT:** Applies a T5 rewriting query model using the full text of the provided canonical responses and all historical utterances to rewrite the current utterance.
- **uogTr-MI-HB:** Applies a T5 rewriting query model using the output from the mixed-initiative sub-task as the context and all historical utterances to rewrite the current utterance.

Provided Baselines

For comparison with our runs' results, we include the results of the baseline methods provided by the track organisers [14], namely:

- **BM25_T5_BART_automatic:** An automatic baseline run that has been provided by the organisers. The top 1000 passages were retrieved and re-ranked using BM25 and a T5-re-ranker using each turn's automatic rewrite. To generate the response answer, this baseline system used a BART model to summarise the top three passages from the retrieval stage.
- **BM25_T5_BART_manual:** A manual baseline run that has been provided by the organisers. The top 1000 documents were retrieved and re-ranked using BM25 and a T5-re-ranker using each turn's manual rewrite. To generate the response answer, this baseline system used a BART model to summarise the top three passages from the retrieval stage.

5 RESULTS

We first report our evaluation results for the Mixed-Initiative sub-task in Section 5.1. The evaluation results for the Conversation Search task are provided in Section 5.2.

5.1 Mixed-Initiative Sub-task Results

Table 2 shows the obtained effectiveness results for our submitted runs and the baseline runs provided by the track organisers in terms of P@1. All evaluation metrics are calculated using the official qrels. Following the guidelines provided by the organisers [14], the performance evaluation is conducted using the P@1 metric (with a threshold of 2), taking the average across 205 users' utterances. This evaluation is based on three criteria: Relevance@1, Novelty@1, and Diversity@1. Diversity measures the number of options provided in the question. Novelty evaluates whether the question adds new information to the conversation. Relevance assesses whether the question logically flows from previous utterances.

From the table, we observe that our run, namely uogTr-MI, achieves the highest performance compared to both the generation and selection baselines on Relevance@1 and Novelty@1. However, the highest performance for Diversity@1 is achieved by the baseline miniLM-BERT. Furthermore, we use the T5QR model (see Section 2.1) to reformulate the current utterance by using clarifying questions from our uogTr-MI run and the user feedback provided by the organisers. This rewritten utterance is then processed through the retrieval pipeline, as outlined in Sections 2.3–2.4, and submitted as the run named uogTr-MI-HB. We present the performance of uog-TR-MI-HB in the next section.

Table 3: Results on the TREC Conversational Assistance Track 2022 Conversational Search task. The best performing run for each measure is emphasised.

Measure	TREC Per-Topic			Baseline		Official Runs		
	Min	Median	Max	automatic	manual	uogTr-AT	uogTr-MI-HB	uogTr-MT
				Lenient				
MAP	0.0180	0.1768	0.4397	0.1628	0.2377	0.1448	0.1679	**0.3391**
NDCG@20	0.0356	0.3203	0.6674	0.3048	0.4333	0.2993	0.3105	**0.4950**
				Strict				
MAP	0.0128	0.1479	0.4265	0.1498	0.2309	0.1343	0.1700	**0.2489**
NDCG@20	0.0356	0.3204	0.6674	0.3048	0.4333	0.2993	0.3143	**0.4639**

5.2 Conversational Search task Results

Table 3 presents the obtained effectiveness results of all our runs in comparison to the provided baselines. The table also shows the TREC per-topic best and median scores across all participating systems, in terms of NDCG@20 and MAP@1000. This year, a run can be evaluated under two relevance thresholds, *lenient*: where passages at least slightly meet the need of the request at that turn (relevance level 1); and *strict*, where passages must at least "moderately meet" the need (relevance level 2). Table 3 (top-half) shows the results of the lenient evaluation, whereas Table 3 (bottom-half) shows the results of the strict evaluation.

Firstly, we analyse the performance of the automatic runs. The results from the table indicate that uogTr-MI-HB outperforms uogTr-AT and BM25_T5_BART_automatic on all measures and under both the lenient and strict evaluation criteria. This is due to the use of the output from the mixed-initiative sub-task as context to rewrite raw utterances. However, our automatic submitted runs (uogTr-MI and uogTr-AT) perform lower than the TREC median. From the table, we also observe that our uogTr-AT run has a lower performance compared to the results of the baseline BM25_T5_BART_automatic. On closer inspection, we observe that the query rewriting model in the baseline uses only the last three (at most) canonical responses as the context while our run uses all of the historical canonical responses. Hence, this might explain the reduced rewriting effectiveness we experienced.

Finally, we analyse the performance of the manual runs. According to the table, our manual run, namely uogTr-MT, outperforms the provided baseline, namely BM25_T5_BART_manual, on all measures and all evaluation criteria (lenient and strict). Moreover, our submitted manual run (uogTr-MT) performs better than the TREC median.

6 CONCLUSIONS

Overall, our participation in the TREC Conversational Assistance track has been valuable in increasing our understanding of effectively leveraging the Multi-Task Learning (MTL) methods to address the Conversational task. We found that our most effective run uogTr-MT using our MTL monoQA model outperform the TREC median on all measures and all evaluation criteria. Moreover, for the mixed-initiative sub-task, our uogTr-MI run – which uses our MTL T5MI model – outperform all of the baselines provided by the organisers on all measures. For future work, we plan to incorporate the query rewriting, retrieval, and reader (answer extractor) components into a single language model for improved efficiency and for simplifying the stages of the pipeline.

Acknowledgements

We would like to thank Dr Sean MacAvaney for his assistance in preparing the format of the runs.

REFERENCES

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). *arXiv preprint arXiv:2009.11352* (2020). https://doi.org/10.48550/arXiv.2009.11352

[2] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Trans. Inf. Syst.* (2002). https://doi.org/10.1145/582415.582416

[3] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proc. EMNLP*. 5918–5924. https://aclanthology.org/D19-1605

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *In Proc. IEEE Transactions on Big Data* (2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[5] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries with a Single QA System. In *Proc. EMNLP*. 1896–1907. https://doi.org/10.18653/v1/2020.findings-emnlp.171

[6] Sarawoot Kongyoung, Craig Macdonald, and Iadh Ounis. 2022. monoQA: Multi-Task Learning of Reranking and Answer Extraction for Open-Retrieval Conversational Question Answering. In *Proc. EMNLP*. 7207–7218. https://aclanthology.org/2022.emnlp-main.485

[7] Tingting Liang, Yixuan Jiang, Congying Xia, Ziqiang Zhao, Yuyu Yin, and Philip S Yu. 2022. Multifaceted Improvements for Conversational Open-Domain Question Answering. *arXiv preprint arXiv:2204.00266* (2022). https://doi.org/10.48550/arXiv.2204.00266

[8] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386* (2020). https://doi.org/10.48550/arXiv.2010.11386

[9] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized Query Embeddings for Conversational Search. In *Proc. EMNLP*. 1004–1015. https://doi.org/10.18653/v1/2021.emnlp-main.77

[10] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proc. RepL4NLP*. 163–173. https://doi.org/10.18653/v1/2021.repl4nlp-1.17

[11] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proc. EMNLP*. https://aclanthology.org/2022.emnlp-main.669

[12] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Proc. EMNLP 2020*. 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[13] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* (2019). https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-v2.pdf

[14] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, and Johanne R. Trippas. 2022. CAsT 2022: TREC CAsT 2022 Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Proc. TREC*.

[15] Chen Qu, Liu Yang, Cen Chen, W Bruce Croft, Kalpesh Krishna, and Mohit Iyyer. 2021. Weakly-Supervised Open-Retrieval Conversational Question Answering.

In *Proc. ECIR*. 529–543. https://doi.org/10.1007/978-3-030-72113-8_35

[16] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proc. SIGIR*. 539–548. https://doi.org/10.1145/3397271.3401110

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Proc. JMLR*. 1–67. http://jmlr.org/papers/v21/20-074.html

[18] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Proc. NIPS*. https://dl.acm.org/doi/pdf/10.5555/3495724.3496209

[19] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2021. University of Glasgow Terrier Team at the TREC 2021 Deep Learning Track.. In *Proc. TREC*. https://trec.nist.gov/pubs/trec30/papers/uogTr-DL.pdf