A Survey on Asking Clarification Questions Datasets in Conversational Systems

Hossein A. Rahmani[†*] **Xi Wang**[†*] **Yue Feng**[†] **Qiang Zhang**[‡] **Emine Yilmaz**[†] **Aldo Lipani**[†]
[†]University College London, London, UK
[‡]Zhejiang University, Hangzhou, China
{hossein.rahmani.22,xi-wang,yue.feng.20,emine.yilmaz,aldo.lipani}@ucl.ac.uk
qiang.zhang.cs@zju.edu.cn

Abstract

The ability to understand a user's underlying needs is critical for conversational systems, especially with limited input from users in a conversation. Thus, in such a domain, Asking Clarification Questions (ACQs) to reveal users' true intent from their queries or utterances arise as an essential task. However, it is noticeable that a key limitation of the existing ACQs studies is their incomparability, from inconsistent use of data, distinct experimental setups and evaluation strategies. Therefore, in this paper, to assist the development of ACQs techniques, we comprehensively analyse the current ACQs research status, which offers a detailed comparison of publicly available datasets, and discusses the applied evaluation metrics, joined with benchmarks for multiple ACQs-related tasks. In particular, given a thorough analysis of the ACQs task, we discuss a number of corresponding research directions for the investigation of ACQs as well as the development of conversational systems.

1 Introduction

Humans often resort to conversations and asking clarification questions to avoid misunderstandings when collaborating with others. Asking Clarification Questions (ACQs) is, therefore, a commonly used mechanism to boost efficiency on human-human as well as human-machine collaborative tasks (Shi et al., 2022; Zou et al., 2023; Shi et al., 2023; Feng et al., 2023). As an example of human-machine collaboration, conversational systems are developed to not only have a natural conversation with people but also to answer various questions of topics ranging from different domains (e.g., news, movie, and music) in an accurate and efficient manner (Gao et al., 2018). To effectively and efficiently answer various questions, it is essential for many existing conversational systems to capture people's intents. Only then can conversational systems accurately reply to a series of questions from users (Anand et al., 2020; Zamani et al., 2022).

Nevertheless, one essential issue is that limited research exists on ACQs and most systems were trained with inconsistent and limited input of data resources. Indeed, in the literature, many studies introduced ACQs to assist conversational systems when applying to different / a mixture of domains (e.g., movie (Li et al., 2017) or open domain (Aliannejadi et al., 2019)). There is also a lack of commonly agreed benchmark datasets for the development of ACQs systems with comparable result analysis. However, on the other hand, in the literature (Aliannejadi et al., 2019; Zamani et al., 2020; Kumar and Black, 2020; Feng et al., 2023), a growing number of studies released publicly available datasets while showing a common interest in the ACQ research direction. This observed contradiction leads to a necessity for a comprehensive overview of the existing datasets as well as the current status of the ACQ research direction. By addressing this concern, many growing ACQs can be better designed, trained and tested with suitable features from properly selected datasets according to comprehensive guidance.

Therefore, in this paper, we offer an overview of the current status of the ACQ research progress. In particular, we aggregate and compare the datasets that have been considered for evaluating recent ACQ techniques from various aspects, such as their dimension, resource, recency and semantic closeness. Afterwards, with the overall discussion of publicly available datasets, we shed light on the model performance while running experiments of corresponding representative techniques on such datasets. Note that, we also release our implementation code for such experiments[1]. Next, we summarised the concluding remarks as well as follow-up suggestions for developing the ACQ techniques.

[*] Equal Contribution

[1] https://github.com/rahmanidashti/ACQSurvey

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2698–2716
July 9-14, 2023 ©2023 Association for Computational Linguistics

Table 1: A statistical summary of ACQ datasets for both Conv. Search and Conv. QA. The highlighted colours indicate the distinct corpus size of datasets (best viewed in colour).

Dataset	# Domains	Scale	# Clar. Q	Link
Conversational Search				
ClariT (Feng et al., 2023)	-	108K	260K	github.com/sweetalyssum/clarit
Qulac (Aliannejadi et al., 2019)	198	10K	3K	github.com/aliannejadi/qulac
ClariQ (Aliannejadi et al., 2021)	300	2M	4K	github.com/aliannejadi/ClariQ
TavakoliCQ (Tavakoli et al., 2021)	3	170K	7K	github.com/Leila-Ta/Clarification_CQA
MIMICS (Zamani et al., 2020)	-	462K	586K	github.com/microsoft/MIMICS
MANtIS (Penha et al., 2019)	14	80K	435	guzpenha.github.io/MANtIS/
ClariQ-FKw (Sekulić et al., 2021)	230	2K	2K	github.com/isekulic/CQ-generation
MSDialog (Qu et al., 2018)	12	35K	877	ciir.cs.umass.edu/downloads/msdialog
MIMICS-Dou (Tavakoli et al., 2022)	-	1K	1K	github.com/Leila-Ta/MIMICS-Duo
Conversational Question Answering				
ClarQ (Kumar and Black, 2020)	173	2M	2M	github.com/vaibhav4595/ClarQ
RaoCQ (Rao and Daumé III, 2018)	3	77K	770K	github.com/raosudha89/ranking_clarification_questions
AmazonCQ (Rao and Daumé III, 2019)	2	24K	179K	github.com/raosudha89/clarification_question_generation_pytorch
CLAQUA (Xu et al., 2019)	110	40K	40K	github.com/msra-nlc/MSParS_V2.0

Our Contributions. The main contributions of this work can be summarized as follows:

- We systematically search through 77 relevant papers, selected as per their recency, reliability and use frequency, in the ACQ domain from top-tier venues.
- We compare the ACQ datasets from their contributions to the development of ACQ techniques and experimentally show the performance of representative techniques.
- We introduce a visualised semantic encoding strategy to explain dataset suitability when selected for their corresponding experiments.
- We analytically outline promising open research directions in the construction of future datasets for ACQs, which sheds light on the development of future research.

2 Conversational Systems

A conversational system functions to assist users while addressing various tasks or acting as a partner in casual conversations (Gao et al., 2018). In particular, conversation systems can be classified into four main categories: (1) Conversational Search (Conv. Search); (2) Conversational Question Answering (Conv. QA); (3) Task-oriented Dialogues Systems (TDSs); and (4) Social Chatbots (Gao et al., 2019; Anand et al., 2020). In particular, the first two types, *Conv. Search* and *Conv. QA*, extend the classic search and QA systems to a conversational nature (Anand et al., 2020; Zaib et al., 2021). For TDSs and social chatbots, they are more recent research topics and were introduced to build systems for assisting users while addressing a specific task or offering emotional connection and companionship via conversations (Gao et al., 2019). However, due to the limited resources that investigate the challenge of asking clarification questions when developing these two systems, this study focuses on Conv. Search and Conv. QA systems.

Moreover, ACQs in conversational systems partially focus on three main tasks, namely, Clarification Need Prediction (T_1), Asking Clarification Questions (T_2), and User Satisfaction with CQs (T_3) (Zamani et al., 2020; Tavakoli et al., 2022; Aliannejadi et al., 2019). First, T_1 evaluates the necessity of asking clarification questions when users provide their initial queries or requests. Next, with a positive decision, we turn to the action of providing suitable clarification questions (i.e., T_2) by following two main routines: generation or selection from a pool of candidate clarification questions. Afterwards, the third task T_3 is to evaluate the effectiveness of the corresponding clarification questions while considering user satisfaction levels from multiple aspects (e.g., the usefulness or relevance of clarification questions). An effective ACQ-encoded conversational system requires a joint effort to address the three tasks satisfactorily to enhance users' conversational experience. Therefore, in this survey, we explore the relevant ACQ datasets and discuss their suitability while addressing the above three tasks.

3 ACQ Datasets

In this section, we describe the main characteristics of the existing and relevant ACQ datasets. Note that we include some additional information, such as the corresponding institution, in Appendix A. A careful dataset selection and aggregation strat-

Table 2: A Summary of collection details of ACQ datasets. '-' means that the information is not available. 'SE' is StackExchange, 'MC' refers to Microsoft Community, and 'KB' is Knowledge Base. The detailed information of each dataset, such as the exact source domains, can be accessed in Appendix A.

Dataset	Published	Built	Resource	Clar. Source
Conversational Search				
ClariT (Feng et al., 2023)	2023	Aug. 2018	General queries from task-oriented dialogues	Crowdsourcing
Qulac (Aliannejadi et al., 2019)	2019	2009-2012	198 topics from TREC WEB Data	Crowdsourcing
ClariQ (Aliannejadi et al., 2021)	2021	2009-2014	300 topics from TREC WEB Data	Crowdsourcing
TavakoliCQ (Tavakoli et al., 2021)	2021	Jul. 2009 to Sep. 2019	3 domains of SE	Post and Comment
MIMICS (Zamani et al., 2020)	2020	Sep. 2019	General queries from Bing users	Machine Generated
MANtIS (Penha et al., 2019)	2019	Mar. 2019	14 domains of SE	Post and Comment
ClariQ-FKw (Sekulić et al., 2021)	2021	2009-2014	TREC WEB Data	Crowdsourcing
MSDialog (Qu et al., 2018)	2018	Nov. 2005 to Oct. 2017	4 domains of MC	Crowdsourcing
MIMICS-Duo (Tavakoli et al., 2022)	2022	Jan. 2022 to Feb. 2022	General queries from Bing users	HIT on MTurk, Qualtrics
Conversational Question Answering				
ClarQ (Kumar and Black, 2020)	2020	-	173 domains of SE	Post and Comment
RaoCQ (Rao and Daumé III, 2018)	2018	-	3 domains of SE	Post and Comment
AmazonCQ (Rao and Daumé III, 2019)	2019	-	A category of Amazon dataset	Review and Comment
CLAQUA (Xu et al., 2019)	2019	-	From an open-domain KB	Crowdsourcing

egy[2] has been applied to this survey to ensure their recency and accessibility.

To offer an overview of dataset dimensions, in Table 1, we describe the ACQ datasets in statistics, together with links to access the datasets. The statistical information includes the number of the considered domains from the corresponding resource; the size of the whole dataset; the number of clarification questions in each dataset. These datasets can be grouped into three sets (large, medium and small, highlighted in pink, cyan and yellow colours) with varied scales of datasets: 1) Large datasets with greater than 10k clarification questions (i.e., ClariT, MIMICS, ClarQ, RaoCQ, AmazonCQ, CLAQUA). Note that all the Conv. QA datasets are classified as large datasets due to the fact that it is more convenient to prepare clarification questions within a QA pair than in a dialogue. 2) Medium datasets with no less than 1K clarification questions (i.e., Qulac, ClariQ, TavakoliCQ, ClariQ-FKw, MIMICS-Dou); 3) Small datasets that have no more than 1K instances and only include MANtIS and MSDialog. In what follows, we compare datasets for developing conversational search and QA systems, according to their key characteristics.

3.1 Conversational Search

Conversational Search (Conv. Search) refers to information retrieval systems that permit a mixed-initiative interaction with one or more users using a conversational interface (Anand et al., 2020). To develop effective Conv. Search systems, many previous studies released a number of datasets and made them publicly available. Here, we briefly describe such datasets:

- **ClariT (Feng et al., 2023):** The first clarification question dataset for task-oriented information seeking, which asks questions to clarify user requests and user profiles based on task knowledge.
- **Qulac (Aliannejadi et al., 2019):** The first clarification question dataset in an open-domain information-seeking conversational search setting with a joint offline evaluation framework.
- **ClariQ (Aliannejadi et al., 2020, 2021):** An extended Qulac with additional crowdsourced topics, questions and answers in the training corpus as well as synthetic multi-turn conversations.
- **TavakoliCQ (Tavakoli et al., 2021; Tavakoli, 2020):** It includes clarification questions collected from the StackExchange QA community and based on three resource categories that have the top number of posts.
- **MIMICS (Zamani et al., 2020):** This dataset comprises three sub-datasets that are all sourced from the application of the clarification pane in Microsoft Bing. In particular, they differ in if such a sub-dataset is based on single or multiple clarification panes (i.e., MIMICS-Click or ClickExplore) or focusing on real search queries and their corresponding query-clarification pairs (i.e., MIMICS-Manual).

[2] We exclude datasets released before 2015 and the ones that are not publicly available.

- **MANtIS (Penha et al., 2019):** A multi-domain (14 domains) conversational information-seeking dataset, sourced from StackExchange, like TavakoliCQ, with joint user intent annotations on the included utterances.

- **ClariQ-FKw (Sekulić et al., 2021):** This dataset introduces facets (the keywords that disambiguate a query) to the ClariQ, which results in an updated version with a set of query-facet-clarification question triples.

- **MSDialog (Qu et al., 2018):** This dataset was constructed from the dialogues on Microsoft Community[3] – a forum that provides technical support for Microsoft products – and also details user intent types on an utterance level.

- **MIMICS-Duo (Tavakoli et al., 2022):** A dataset, stands upon the queries from MIMICS-ClickExplore, that enables both online and offline evaluations for clarification selection and generation approach.

3.2 Conversational Question Answering

The idea behind Conversational Question Answering (Conv. QA) is to ask the system a question about a provided passage offering a conversational interface (Zaib et al., 2021). Conv. QA has recently received growing attention in the research community while introducing multiple available large-scale datasets. A brief discussion of such datasets are as follows:

- **ClarQ (Kumar and Black, 2020):** This dataset is sourced from the post-question pairs in StackExchange and developed with self-supervised approaches within a bootstrapping framework.

- **RaoCQ (Rao and Daumé III, 2018):** Another StackExchange-based dataset with a large volume of post-question-answer triples from three selected domains.

- **AmazonCQ (Rao and Daumé III, 2019):** An Amazon platform-based Clarification QA dataset with questions targeting the missing information of products and answers provided by sellers or other users. In addition, a context is offered that contains both the product title and description.

Table 3: Summary of tasks and evaluation method on ACQs datasets. The tasks can be generation and ranking, which are indicated by 'G' and 'R', respectively.

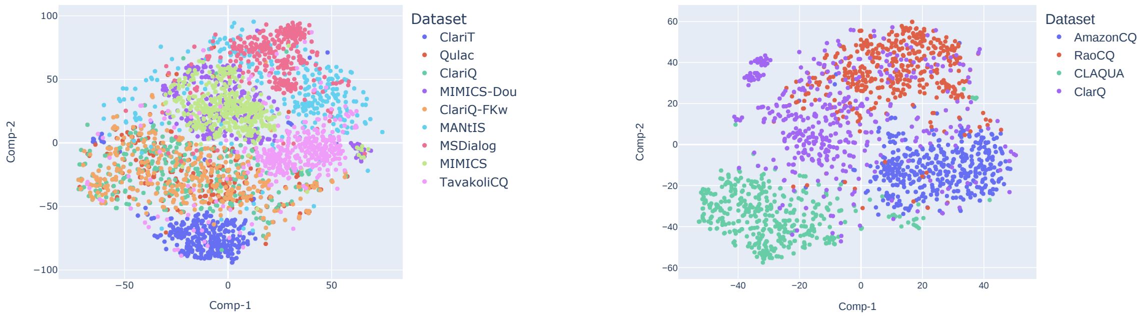
Dataset	Task			Eval. Method
	T_1	T_2	T_3	
		Conv. Search		
ClariT (2023)	✓	G	-	Offline
Qulac (2019)	-	R	-	Offline
ClariQ (2021)	✓	R	-	Offline
TavakoliCQ (2021)	-	G	-	Offline
MIMICS (2020)	✓	R, G	✓	Offline/Online
MANtIS (2019)	-	R, G	-	Offline
ClariQ-FKw (2021)	-	G	-	Offline
MSDialog (2018)	-	R, G	-	Offline
MIMICS-Duo (2022)	✓	R, G	✓	Offline/Online
		Conv. QA		
ClarQ (2020)	-	R	-	Offline
RaoCQ (2018)	-	R	-	Offline
AmazonCQ (2019)	-	G	-	Offline
CLAQUA (2019)	✓	G	-	Offline

- **CLAQUA (Xu et al., 2019):** A clarification-focus dataset that supports the supervised evaluation of text understanding and generation modules, along with a knowledge-based QA system (KBQA).

3.3 Datasets Analysis

As discussed in Section 1, a major concern of developing the techniques for asking clarification questions is using suitable datasets to train, validate and test the corresponding approach. In particular, it is essential to be aware of the information on when, how and where a dataset is collected. Such information offers a comprehensive description of datasets for their various characteristics, such as their recency and reliability. Therefore, in Table 2, we describe the collection details of each ACQ dataset. In particular, we include the time when the datasets were built as well as the year the corresponding papers were published to indicate the recency of the datasets. In addition, we summarise the source of the data collection, which tells where the datasets came from. Next, we aggregate the main strategies for preparing the clarification questions. At first, due to our data selection strategy, most of the datasets are based on relatively recent information. However, we still observe that some datasets rely on the data collected years ago. For example, the Qulac, ClariQ and ClariQ-FKw datasets consistently use the TREC WEB data but run between 2009 and 2014. The most recent dataset is MIMICS-Duo which was built in 2022, and ClariT is the most recently published dataset in 2023. In particular, all the Conv. QA datasets are limited,

[3] https://answers.microsoft.com/



(a) tSNE on Conv. Search Datasets

(b) tSNE on Conv. QA Datasets

Figure 1: tSNE on ACQ Datasets

with no time information on when their data was collected, which makes them incomparable based on this measure. On the other hand, regarding how and where the datasets were collected, the TREC WEB data, StackExchange and Bing are the commonly considered resource for preparing clarification questions in a dataset. Such platforms' search and question-answering nature is the leading cause of such a finding. Afterwards, the crowdsourcing strategy is commonly applied to generate qualified clarification questions. Note that the posts and comments of StackExchange are also widely used to provide clarification questions. According to the provided information, we conclude that the datasets have been collected based on varied strategies, on different periods and use inconsistent resources. However, it is difficult to tell how exactly a dataset is different from others and how to properly select a set of datasets to show the performance of a newly introduced model. Therefore, in this survey, we introduce a visualisation-based approach to assist the selection of datasets for an improved experimental setup.

In Figures 1a and 1b, we use the t-distributed Stochastic Neighbor Embedding (i.e., t-SNE) method to visualize the semantic representation of clarification questions (semantic embeddings) for Conv. Search and Conv. QA datasets. As one can see from Figure 1a, Qulac and ClariQ datasets, and MIMICS and MIMICS-Dou datasets highly overlapped with each other. It was expected to be seen as ClariQ and MIMICS-Duo are built on top of Qulac and MIMICS, respectively. This indicates that achieving a high-quality performance of a proposed asking clarification model on both Qulac and ClariQ (or MIMICS and MIMICS-Duo) is not satisfactory as they include clarification questions with close semantic meanings. Figure 1a shows that Conv. Search datasets form 5 distinct clusters that can be used to evaluate asking clarification models. For example, the models' generalisability can be evaluated on the ClariT, Qulac, TavakaliCQ, MIMICS, and MSDialog datasets, which locates with few overlapped instances between them. More importantly, comparing Figures 1a and 1b reveals that clarification questions in Conv. Search are very focused while the clarification questions in Conv. QA datasets are more widely distributed. This indicates the high similarities among the Conv. Search-based data and the resulting necessity of properly selecting those publicly available datasets.

4 Evaluation Metrics

In this section, we detail the description of the applicable evaluation metrics for the included datasets when evaluating ACQs approaches. In particular, as previously discussed, we discuss such metrics accordingly if they are automatic or human-involved.

4.1 Automatic Evaluation

With a ready dataset, ACQ-based conversational systems can be evaluated using a variety of automatic evaluation metrics. The widely-used metrics can be categorized into two groups based on the strategy of giving clarification questions, i.e., ranking or generation. For the ranking route, the commonly used evaluation metrics include (1) MAP (Jarvelin, 2000), (2) Precision (Järvelin and Kekäläinen, 2017), (3) Recall (Jarvelin, 2000), (4) F1-score (Beitzel, 2006), (5) Normalized Discounted Cumulative Gain (nDCG) (Wang et al., 2013), (6) Mean Reciprocal

Rank (MRR) (Voorhees et al., 1999; Radev et al., 2002), and (7) Mean Square Error (MSE) (Beitzel, 2006). The main idea behind using these metrics is to evaluate the relevance of the top-ranked clarification questions by the system to reveal the corresponding user intent. On the other hand, some common metrics for the generation route include (8) BLEU (Papineni et al., 2002), (9) METEOR (Banerjee and Lavie, 2005), (10) ROUGE (Lin, 2004). BLEU and ROUGE were originally developed to evaluate machine translation and text summarization results, respectively. Recently, they have also been applied as evaluation metrics while addressing the ACQ task (Sekulić et al., 2021; Zhang and Zhu, 2021; Shao et al., 2022). Their scores are both based on the n-gram overlap between generated and reference questions. The difference between BLEU and ROUGE corresponds to the precision and recall metrics. BLEU calculates the ratio of predicted terms in the reference question, while ROUGE scores indicate the ratios of terms from the reference are included in the predicted text. Next, ROUGE-L, a newer version of ROUGE – focuses on the longest common subsequence – is recently being used in evaluating ACQ models. However, these above metrics are limited while ignoring human judgements. Therefore the METEOR was introduced to address such a concern by considering the stems, WordNet synonyms, and paraphrases of n-grams.

The main advantage of using automatic evaluation metrics is that they are not expensive for consideration and can be applied easily. However, they are not always aligned with human judgments. Therefore, recent studies also consider human evaluation besides their automatic evaluation to show how the generated or selected CQs impact on the performance of their conversation systems.

4.2 Human Evaluation

In addition to automatic evaluation metrics, human evaluation provides a more accurate and qualitative evaluation of generated or ranked CQs. An essential reason is that automatic evaluation metrics mainly consider n-gram overlaps or ranking of CQs instead of their semantic meaning or other quality-wise aspects. Thus, human annotations are increasingly used to evaluate clarifying questions. The human annotation process consists of scoring generated or selected CQs based on several quality dimensions. Compared to automatic evaluation, human evaluation is naturally more expensive due to the manual annotation effort, but it provides a more accurate picture of the quality of the output. The main aspects that are evaluated using human annotations include (1) *relevance* (Aliannejadi et al., 2020), which shows if a CQ is relevant to the user's information need (2) *usefulness* (Rosset et al., 2020) that is related to adequacy and informativeness of a question, (3) *naturalness* (Li et al., 2019) that evaluates a question if it is natural, fluent, and likely generated by a human and (4) *clarification* (Aliannejadi et al., 2021) that shows how the user's feedback influences the model's next CQ question. There are also *humanness* (See et al., 2019), *engangingness* (Li et al., 2019), *interestingness* (Li et al., 2019), *knowledgeable* (Li et al., 2019), that evaluate a CQ by considering the whole conversation, instead of an individual query-question pair. However, the ACQ domain lacks a consistent or agreed terminology for the used human evaluation metrics. In addition, some of them could have overlapped focus when evaluating the clarification questions. For example, the *usefulness* can also be evaluated based on the *knowledgeable* of the corresponding clarification question.

Table 4: Clarification need prediction performance of best representative methods from traditional ML and language models (RandomForest and BERT) on datasets. ↑ or ↓ is added to BERT to indicate a consistent performance change on all evaluation metrics. (The results of all methods are added to Table 7 in Appendix B.1).

Model	**Precision**	**Recall**	**F1**
		ClariQ	
RandomForest	0.3540	**0.3806**	**0.3717**
BERT	**0.3804**	0.3249	0.3344
		CLAQUA	
RandomForest	0.2860	0.5000	0.3638
BERT ↑	**0.6349**	**0.625**	**0.6255**
Model	**MAE**	**MSE**	**R^2**
		MIMICS	
RandomForest	**2.4404**	**7.969**	**-0.0012**
BERT ↓	2.4562	8.1277	-0.0211
		MIMICS-Duo	
RandomForest	**2.8502**	**11.206**	**-0.0079**
BERT ↓	2.8801	11.2268	-0.0098

5 Model Performance on ACQ

In this section, to offer a complete view of the current progress of the ACQ task, we discuss the main observations of the recent ACQ techniques when running on various ACQ datasets. Moreover, for each of the ACQ-related tasks, i.e., T_1, T_2 and T_3, we show the performance of many commonly used baselines while running on the applicable datasets for offering some additional concluding remarks.

Table 5: Question relevance ranking performance evaluation on representative approaches. 'P' and 'R' refers to Precision and Recall. ↑ or ↓ is added to Doc2Query + BM25 to indicate a consistent performance change to BM25 on all evaluation metrics.

Model	**MAP**	**P@10**	**R@10**	**NDCG**
	Qulac			
BM25	**0.6306**	**0.9196**	0.1864	0.9043
Doc2Query + BM25	0.6289	**0.9196**	0.1860	**0.9069**
	ClariQ			
BM25	0.6360	0.7500	0.5742	0.7211
Doc2Query + BM25 ↑	**0.6705**	**0.7899**	**0.6006**	**0.7501**
	TavakoliCQ			
BM25	0.3340	0.0463	0.4636	0.3743
Doc2Query + BM25 ↑	**0.3781**	**0.0540**	**0.5405**	**0.4260**
	MANtIS			
BM25	0.6502	0.0679	0.6795	0.6582
Doc2Query + BM25 ↑	**0.7634**	**0.0830**	**0.8301**	**0.7802**
	ClariQ-FKw			
BM25	**0.7127**	0.5880	0.7181	**0.7910**
Doc2Query + BM25	0.7073	**0.5940**	**0.7244**	0.7874
	MSDialog			
BM25	**0.8595**	**0.0929**	**0.9293**	**0.8781**
Doc2Query + BM25 ↓	0.8430	0.0908	0.9087	0.8624
	ClarQ			
BM25	**0.2011**	**0.0259**	**0.2596**	**0.2200**
Doc2Query + BM25 ↓	0.1977	0.0263	0.2630	0.2168
	RaoCQ			
BM25	**0.1511**	0.0236	0.2362	0.1797
Doc2Query + BM25	0.1509	**0.0241**	**0.2415**	**0.1811**
	CLAQUA			
BM25	**0.9600**	**0.0992**	**0.9920**	**0.9683**
Doc2Query + BM25 ↓	0.9395	0.0990	0.9901	0.9523

First, according to our exploration of experimental results of recent ACQ techniques, we observe three main limitations of their inconsistent experimental setups, used baselines and model generalisability. Indeed, many research studies have inconsistent uses of datasets as well as incomparable results with distinct experimental setups. For example, Krasakis et al. (2020) and Bi et al. (2021) both used the Qulac dataset. In (Krasakis et al., 2020), they randomly kept 40 topics for testing their performance of a heuristic ranker. However, instead of following (Krasakis et al., 2020), Bi et al. (2021) used a few-turn-based setup while leveraging the Qulac dataset for asking clarification questions. Next, another common issue is the use of different baselines to show the leading performance of newly introduced techniques. For example, the study in (Aliannejadi et al., 2019) primarily employed ranking-based models, such as RM3, LambdaMART, and RankNet, to evaluate the performance of their question retrieval model. In contrast, the study in (Aliannejadi et al., 2021) utilized language models like RoBERTa and ELECTRA to evaluate the performance of their question relevance model. More importantly, many techniques were introduced while tested on a single dataset to show their top performance (e.g., (Krasakis et al., 2020; Sekulić et al., 2022; Zhao et al., 2022)), which lead to a significant generalisability concern. This also indicates the necessity of developing a benchmark while evaluating the ACQ techniques and identifying the exact state-of-the-art. Next, to acquire an overview of model performance while running experiments on the included datasets, we present the experimental results with representative approaches on the three ACQs subtasks, i.e., T_1, T_2 and T_3 that are discussed in Section 2. The details of our experiments can be found in Appendix B. Table 4 shows the results of two top-performing models (i.e., BERT and RandomForest) for the clarification need prediction task (T_1) from traditional ML and language models. A key observation is that the prediction of clarification need should be selectively made in a classification or regression setup. In particular, BERT, a language model that well classifies the classification need on ClariQ and CLAQUA datasets, does not consistently outperform a classic approach, RandomForest, in addressing a regression-wise task (as per the results on MIMICS and MIMICS-Duo). Next, for the second sub-task, ask clarification questions, which can be addressed via generation or ranking. However, clarification question generation requires a detailed context description and associated information. The existing approaches (e.g., Seq2Seq models) could be either naive in solely taking the query as input for CQ generation or difficult to generalise to many datasets while using specific information. Therefore, in this study, we compare the ranking performance when applying some commonly used ranking baselines (i.e., BM25 and BM25 with query expanded via the Doc2Query technique (Nogueira et al., 2019)) on every dataset. Table 5 presents the experimental results of these two approaches on every dataset. Note that, we ignore the experimental results on ClariT, MIMICS, MIMICS-DUO and AmazonCQ since they are different from other datasets in having queries with multiple relevant clarification questions. For the results, we observe that the query expansion via Doc2Query can be effective for most of the conversational search datasets, due to their shorter queries. However, when query expansion is applied to a Conv. QA dataset, it is not promising for an improved performance. Another observation is that the Qulac, ClariQ and ClariQ-FKw datasets have similar clarification questions in their dataset as per Figure 1a and Doc2Query-based query expansion has limited improvement to BM25 on these datasets. However, for another two corpus, TavakoliCQ and MANtIS, with distinct clarification questions, a bigger improvement margin can be observed. This also indicates the usefulness of our introduced visualisation-based strategy for dataset selection.

Next, for the third task, it is crucial to determine user satisfaction with clarification questions (CQs), as it provides insight into how well the CQs are serving their intended purpose. However, obtaining the necessary data for evaluating user satisfaction can be challenging. In the literature, only two datasets (i.e., MIMICS and MIMICS-Duo) include information for this task. In Table 6, we present the corresponding results. A similar observation to the clarification need prediction task is that the language model can assist an ACQ technique in effectively evaluating user satisfaction. However, due to the limited number of applicable datasets, this observation might not be consistent in a different context. This also aligns with the current status of the ACQ research task while evaluating the newly proposed ACQ techniques.

Overall speaking, with the presented experimental results, we indicate the inconsistent performance of models while evaluated on different datasets. In particular, we also discuss the limited numbers of useful datasets while evaluating ACQ techniques (e.g., the models' performance on user satisfaction prediction).

6 Discussion and Future Challenges

From the exploration of datasets as well as the experimental results on them, in this section, we highlight the concluding remarks on the current status of the ACQ research task, mainly from the dataset point of view. In addition, we discuss the promising directions based on the main findings listed below.

Table 6: User satisfaction prediction with CQs performance of running best representative methods from traditional ML and language models (MultinomialNB and distilBERT) on datasets. ↑ is added to distilBERT to indicate a consistent performance improvement on all evaluation metrics. (The results of all methods are added on Table 8 in Appendix B.3).

Model	Precision	Recall	F1
	MIMICS		
MultinomialNB	0.8255	0.7842	0.7758
distilBERT ↑	**0.9453**	**0.9397**	**0.939**
	MIMICS-Duo		
MultinomialNB	**0.4407**	0.2787	0.2336
distilBERT	0.2766	**0.2803**	**0.2777**

Findings. (1) **Missing Standard Benchmark.** Existing datasets are underdeveloped, and difficult to constitute a standard benchmark while introducing novel ACQ techniques. As a consequence, it is challenging to effectively and accurately compare the proposed techniques and capture the true state-of-the-art. (2) **Few User-System Interactions Recorded for Evaluation.** In the literature, only the MIMICS dataset was collected by using a clarification pane that simulates such interactions. This makes it challenging to evaluate models in a near-realistic scenario and to estimate how well they could perform in a real-world setting. (3) **Inconsistent Dataset Collection and Formatting.** Many included datasets in this paper are frequently presented in distinct structures and can only be applied with a tailored setup. This is a problem while developing techniques and evaluating them on multiple datasets. (4) **Inconsistent Model Evaluation.** Many newly introduced models apply customised evaluation strategies even while using an identical dataset for addressing a specific asking clarification task. This lead to difficulties in model performance comparison.

Future Research Directions. (1) **Benchmark Development.** For the development of an ACQs technique, it is important that the models are compared to a common-accepted benchmark to make the corresponding conclusions. However, according to the above findings, currently, it is still unavailable. Therefore, benchmark development is the first key future direction. (2) **ACQ Evaluation Framework.** Aside from the benchmark development, it is also essential for a proper evaluation of newly introduced techniques. In particular, due to the human-machine interaction nature of the ACQ techniques, it is valuable for evaluation metrics to take user satisfaction information into account. In addition, the introduction of a corresponding evaluation framework can assist the development of ACQ techniques with systematic evaluations. (3) ***Large-Scale Human-to-Machine Dataset.*** Existing datasets have many limitations that increase the difficulty of developing large-scale models for generating or ranking clarification questions. It remains challenging to collect and build large amounts of data. In the near future, researchers should optimize the process of ACQs based on the current retrieval technologies (see (Trippas et al., 2018) for a description of collecting such datasets). (4) ***Multi-Modal ACQs Dataset.*** Recently multi-modal conversational information seeking has received attention in conversational systems (Deldjoo et al., 2021). Amazon Alexa[4] organised the first conversational system challenge to incorporate multi-modal (voice and vision) customer experience. However, there is a lack of existing datasets containing multi-modal information for ACQs.

Limitations

In this section, we outline the key limitations of our research. Our findings on the ACQ models are not as advanced as the current state-of-the-art, but they serve as a benchmark for others to compare with when using similar datasets. Additionally, to conduct more extensive experiments on larger datasets and more advanced models, we require additional computational resources. Specifically, generating clarification questions is a demanding task as it requires the use of powerful language models.

Acknowledgments

This research is supported by the Engineering and Physical Sciences Research Council [EP/S021566/1] and the EPSRC Fellowship titled “Task Based Information Retrieval” [EP/P024289/1].

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

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*.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR ’19.

Giambattista Amati, Giuseppe Amodeo, Marco Bianchi, Carlo Gaibisso, and Giorgio Gambosi. 2008. Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. Technical report, FONDAZIONE UGO BORDONI ROME (ITALY).

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational search (dagstuhl seminar 19461). In *Dagstuhl Reports*, volume 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Steven M Beitzel. 2006. *On understanding and classifying web queries*. Illinois Institute of Technology.

Keping Bi, Qingyao Ai, and W Bruce Croft. 2021. Asking clarifying questions based on negative feedback in conversational search. In *Proc. of ICTIR*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

[4] `https://www.amazon.science/alexa-prize/taskbot-challenge`

Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multi-modal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*, pages 1577–1587.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yue Feng, Hossein A Rahmani, Aldo Lipani, and Emine Yilmaz. 2023. Towards asking clarification questions for information seeking on task-oriented dialogues. *arXiv preprint arXiv:2305.13690*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations and Trends.

Kalervo Jarvelin. 2000. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2000*.

Kalervo Järvelin and Jaana Kekäläinen. 2017. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In *Proc. of ICTIR*.

Vaibhav Kumar and Alan W Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue learning with human-in-the-loop. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Wei-Yin Loh. 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 989–992.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*. Citeseer.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746.

Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–175.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Exploiting document-based features for clarification in conversational search. In *European Conference on Information Retrieval*.

Taihua Shao, Fei Cai, Wanyu Chen, and Honghui Chen. 2022. Self-supervised clarification question generation for ambiguous multi-turn conversation. *Information Sciences*, 587:626–641.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute or ask clarification questions. *arXiv preprint arXiv:2204.08373*.

Zhengxiang Shi, Jerome Ramos, To Eun Kim, Xi Wang, Hossein A Rahmani, and Aldo Lipani. 2023. When and what to ask through world states and text instructions: Iglu nlp challenge solution. *arXiv preprint arXiv:2305.05754*.

Leila Tavakoli. 2020. Generating clarifying questions in conversational search systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3253–3256.

Leila Tavakoli, Johanne R Trippas, Hamed Zamani, Falk Scholer, and Mark Sanderson. 2022. Mimics-duo: Offline & online evaluation of search clarification. *arXiv preprint arXiv:2206.04417*.

Leila Tavakoli, Hamed Zamani, Falk Scholer, William Bruce Croft, and Mark Sanderson. 2021. Analyzing clarification in asynchronous information-seeking conversations. *Journal of the Association for Information Science and Technology*.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 conference on human information interaction & retrieval*, pages 32–41.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629.

Xin Yan and Xiaogang Su. 2009. *Linear regression analysis: theory and computing*. world scientific.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *arXiv preprint arXiv:2106.00874*.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th acm international conference on information & knowledge management*, pages 3189–3196.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808*.

Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*, pages 3501–3511.

Ziliang Zhao, Zhicheng Dou, Jiaxin Mao, and Ji-Rong Wen. 2022. Generating clarifying questions with web search results. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jie Zou, Aixin Sun, Cheng Long, Mohammad Aliannejadi, and Evangelos Kanoulas. 2023. Asking clarifying questions: To benefit or to disturb users in web search? *Information Processing & Management*, 60(2):103176.

A Datasets Details

A.0.1 ClariT

The ClariT dataset (Feng et al., 2023) was released in 2023 by researchers from the University College London. ClariT is the first dataset for asking clarification questions in task-oriented conversational information seeking. They built ClariT based on an existing dataset ShARC[5], which clarifies users' information needs in task-oriented dialogues. They extended dialogues in ShARC with user profiles to ask clarification questions considering personalized information. To ask clarification questions efficiently, they also removed unnecessary clarification questions in the original dialogues. The collected dataset consists of over $108k$ multi-turn conversations including clarification questions, user profiles, and corresponding task knowledge in general domains.

A.0.2 Qulac

The Qulac (**Qu**estions for **la**ck of **c**arity) (Aliannejadi et al., 2019) dataset is a joint effort by researchers from the Università della Svizzera Italiana and the University of Massachusetts Amherst. Qulac is the first dataset as well as an offline evaluation framework for studying clarification questions in open-domain information-seeking conversational search systems. To acquire the clarification questions, they proposed a four-step strategy: (1) they defined the topics and their facets borrowed from TREC Web Track[6]; (2) they collected several candidates clarification questions for each query through crowdsourcing in which they asked human annotators to generate questions for a given query according to the results showed using a commercial search engine; (3) they assessed the relevance of the questions to each facet and collected new questions for those facets that require more specific questions; (4) finally, they collected the answers for every query-facet-question triplet. The collected dataset consists of over $10,277$ single-turn conversations including clarification questions and their answers on multi-faceted and ambiguous queries for 198 topics with 762 facets.

A.0.3 ClariQ

The ClariQ dataset (Aliannejadi et al., 2020, 2021) was released in 2020 by researchers from the University of Amsterdam, Microsoft, Google, Univer-

[5] https://sharc-data.github.io

[6] https://trec.nist.gov/data/webmain.html

sity of Glasgow, and MIPT. The ClariQ dataset was collected as part of the ConvAI3[7] challenge which was co-organized with the SCAI[8] workshop. The ClariQ dataset is an extended version of Qulac, i.e., new topics, questions, and answers have been added in the training set using crowdsourcing. Like Qulac, ClariQ consists of single-turn conversations (initial_request, followed by clarification questions and answers). Moreover, it comes with synthetic multi-turn conversations (up to three turns). ClariQ features approximately $18K$ single-turn conversations, as well as 1.8 million multi-turn conversations.

A.0.4 TavakoliCQ

Recently Tavakoli et al. (Tavakoli et al., 2021; Tavakoli, 2020), from RMIT University and the University of Massachusetts Amherst, explore the ACQs to provide insightful analysis into how they are used to disambiguate the user ambiguous request and information needs. To this purpose, they extracted a set of clarification questions from posts on the StackExchange question answering community (Tavakoli, 2020). They investigate three sites with the highest number of posts from three different categories covering a period from July 2009 to September 2019. Therefore, the created dataset includes three domains, i.e., business domain with $13,187$ posts, culture with $107,266$ posts, and life/arts with $55,959$ posts. To identify the potential clarification questions, they collected the comments of each post that contain at least one sentence with a question mark, excluding questions submitted by the author of the post and questions that appeared in quotation marks. Their finding indicates that the most useful clarification questions have similar patterns, regardless of the domain.

A.0.5 MIMICS

MIMICS (stands for the **MI**crosoft's **M**ixed-**I**nitiative **C**onversation **S**earch Data) (Zamani et al., 2020). This is a large-scale dataset for search clarification which is introduced in 2020 by researchers from Microsoft. Recently, Microsoft Bing added a clarification pane to its results page to clarify faceted and ambiguous queries.[9] Each clarification pane includes a clarification question and up to five candidate answers. They used internal algorithms and machine learning models based on users' history with the search engine and content analysis to generate clarification questions and candidate answers. The final MIMICS dataset contains three datasets: (1) MIMICS-Click includes $414,362$ unique queries, each related to exactly one clarification pane, and the corresponding aggregated user interaction clicks; (2) MIMICS-ClickExplore contains the aggregated user interaction signals for over $64,007$ unique queries, each with multiple clarification panes, i.e., $168,921$ query-clarification pairs; (3) MIMICS-Manual includes over 2k unique real search queries and 2.8k query-clarification pairs. Each query-clarification pair in this dataset has been manually labeled by at least three trained annotators and the majority voting has been used to aggregate annotations. It also contains graded quality labels for each clarification question, the candidate answer set, and the landing result page for each candidate answer.

A.0.6 MANtIS

The MANtIS (short for **M**ulti-dom**A**i**N** **I**nformation **S**eeking dialogues) dataset (Penha et al., 2019) is a large-scale dataset containing multi-domain and grounded information-seeking dialogues introduced by researchers from TU Delft. They built the MANtIS dataset using extraction of conversations from the StackExchange question answering community. This dataset includes 14 domains on StackExchange. Each question-answering thread of a StackExchange site is a conversation between an information seeker and an information provider. These conversations are included if (1) it takes place between exactly two users; (2) it consists of at least 2 utterances per user; (3) it has not been marked as spam, offensive, edited, or deprecated; (4) the provider's utterances contain at least a reference (a hyperlink), and; (5) the final utterance belongs to the seeker and contains positive feedback. The final MANtIS dataset includes 80k conversations over 14 domains. Then, to indicate the type of user intent, they sampled $1,365$ conversations from MANtIS and annotate their utterances according to the user intent, such as *original question*, *follow-up question*, *potential answer*, *positive feedback*, *negative feedback*, etc. The final sample contains $6,701$ user intent labels.

A.0.7 ClariQ-FKw

The ClariQ-FKw (FKw stands for Facet Keywords) (Sekulić et al., 2021) was proposed by researchers

[7] http://convai.io

[8] https://scai-workshop.github.io/2020/

[9] However, this feature is not yet available for some international markets.

from the University of Amsterdam and the Università della Svizzera Italiana in 2021. Their main objective was to use text generation-based large-scale language models to generate clarification questions for ambiguous queries and their facets, where by facets they mean keywords that disambiguate the query. The dataset includes queries, facets, and clarification questions, which form triplets construed on top of the ClariQ (Aliannejadi et al., 2020) dataset. To this end, they perform a simple data filtering to convert ClariQ data samples to the appropriate triplets and derive the facets from topic descriptions. The final ClariQ-FKw contains $2,181$ triplets.

A.0.8 MSDialog

The MSDialog (Qu et al., 2018) proposed by researchers from the University of Massachusetts Amherst, RMIT University, Rutgers University, and Alibaba Group, is used to analyse information-seeking conversations by user intent distribution, co-occurrence, and flow patterns in conversational search systems. The MSDialog dataset is constructed based on the question-answering interactions between information seekers and providers on the online forum for Microsoft products. Thus, to create the MSDialog dataset, they first crawled over 35k multi-turn QA threads (i.e., dialogues) containing 300k utterances from the Microsoft Community[10] – a forum that provides technical support for Microsoft products – and then annotated the user intent types on an utterance level based on crowdsourcing using Amazon Mechanical Turk (MTurk)[11]. To provide a high-quality and consistent dataset, they selected about 2.4k dialogues based on four criteria, conversations 1) with 3 to 10 turns; 2) with 2 to 4 participants; 3) with at least one correct answer selected by the community, and; 4) that fall into one of the following categories: Windows, Office, Bing, and Skype, which are the major categories of Microsoft products. The final annotated dataset contains $2,199$ multi-turn dialogues with $10,020$ utterances.

A.0.9 MIMICS-Duo

The MIMICS-Duo (Tavakoli et al., 2022) dataset is proposed by researchers at RMIT University, the University of Melbourne, and the University of Massachusetts Amherst. It provides the online and offline evaluation of clarification selection and generation approaches. It is constructed based on the queries in MIMICS-ClickExplore (Zamani et al., 2020), a sub-dataset of MIMICS (Zamani et al., 2020) that consists of online signals, such as user engagement based on click-through rate. The MIMICS-Duo contains over 300 search queries and $1,034$ query-clarification pairs.

A.0.10 ClarQ

The ClarQ dataset (Kumar and Black, 2020) was created in 2020 by Carnegie Mellon University. The ClarQ is designed for large-scale clarification question generation models. To do this, the ClarQ dataset is built with a bootstrapping framework based on self supervision approaches on top of the post-comment tuples extracted from StackExchange[12] question answering community. To construct the ClarQ, they first extracted the posts and their comments from 173 domains. Then, they filtered unanswered posts and only considered comments to posts with at least one final answer as a potential candidate for a clarification question. The ClarQ dataset consists of about 2 million post-question tuples across 173 domains.

A.0.11 RaoCQ

Rao and Daumé III [2018] from the University of Maryland study the problem of ranking clarification questions and propose an ACQs dataset on top of StackExchange. To create this dataset, they use a dump of StackExchange and create a number of post-question-answer triplets, where the post is the initial unedited request, the question is the first comment containing a question (i.e., indicated by a question mark), and the answer is either the edits made to the post after the question (i.e., the edit closest in time following the question) or the author's answer of the post to the question in the comment section. The final dataset includes a total of $77,097$ triples across three domains *askubuntu*, *unix*, and *superuser*.

A.0.12 AmazonCQ

Rao and Daumé III [2019] from Microsoft and the University of Maryland, released a dataset for generating clarification questions. The dataset contains a context that is a combination of product title and description from the Amazon website,a question that is a clarification question asked to the product about some missing information in the context, and the answer that is the seller's (or other users')

[10] https://answers.microsoft.com/

[11] https://www.mturk.com/

[12] https://stackexchange.com/

reply to the question. To construct this dataset, they combined the Amazon Question Answering dataset created by (McAuley and Yang, 2016) and the Amazon Review dataset proposed by (McAuley et al., 2015). The final dataset consists of $15,859$ contexts (i.e., product description) with 3 to 10 clarification questions, on average 7, per context.

A.0.13 CLAQUA

The CLAQUA dataset (Xu et al., 2019) was created by researchers from of Peking University, the University of Science and Technology of China, and Microsoft Research Asia in 2019. They propose the CLAQUA dataset to provide a supervised resources for training, evaluation and creating powerful models for clarification-related text understanding and generation in knowledge-based question answering (KBQA) systems. The CLAQUA dataset is constructed in three steps, (1) sub-graph extraction, (2) ambiguous question annotation, and (3) clarification question annotation. In the first step, they extract ambiguous sub-graphs from an open-domain knowledge base, like FreeBase. They focus on shared-name ambiguity where two entities have the same name and there is a lack of necessary distinguishing information. Then, in the second step, they provide a table listing the shared entity names, their types, and their descriptions. Based on this table, annotators need to write ambiguous questions. Finally, in the third step, based on entities and the annotated ambiguous question, annotators are required to summarize distinguishing information and write a multi-choice clarification question including a spacial character that separate entity and pattern information. They provided these steps for single- and multi-turn conversations. The final CLAQUA dataset contains $17,163$ and $22,213$ single-turn and multi-turn conversations, respectively.

B Experiments on Model Performance

B.1 Clarification Need Prediction

The clarification need prediction is a major task in search clarification to decide whether to ask clarification questions. Between the discussed CQ datasets only ClariQ (Aliannejadi et al., 2020, 2021), MIMICS (Zamani et al., 2020), MIMICS-Duo (Tavakoli et al., 2022), and CLAQUA (Xu et al., 2019) provide the necessary information for the clarification need prediction task. The ClariQ and CLAQUA datasets model the clarification need prediction task as a classification problem. They both present the initial user request with a classification label that indicates the level of clarification required. In contrast to the ClariQ and CLAQUA datasets, the task in the MIMICS and MIMICS-Dou datasets is modelled as a regression task for predicting user engagement. Specifically, these datasets aim to predict the degree to which users find the clarification process useful and enjoy interacting with it. Based on this prediction, the system can make a decision on whether or not to request clarification. We subsequently evaluated the prediction task for clarification needs using a variety of traditional machine learning models and language models. The traditional machine learning models employed as baselines include Random Forest (Breiman, 2001), Decision Tree (Loh, 2011), Multinomial Naive Bayes (MultinomialNB) (Manning, 2008), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and Linear Regression (Yan and Su, 2009). The language model baselines utilized include BART (Lewis et al., 2019), XLNet (Yang et al., 2019), XLM (Lample and Conneau, 2019), Albert (Lan et al., 2019), distilBERT (Sanh et al., 2019), and BERT (Devlin et al., 2018). These models were applied to both classification and regression tasks. The input to traditional ML models is a matrix of TF-IDF features extracted from the raw input text. We use Scikit-learn[13] (Pedregosa et al., 2011), HuggingFace[14] (Wolf et al., 2019), and TensorFlow (Abadi et al., 2016) for the implementation of the aforementioned models.

B.2 Question Relevance Ranking Baselines

To address the second task, namely asking clarification questions, many studies have explored either generation or ranking strategies. However, as we argued in Section 5, the generation techniques require rich information for satisfactory performance and they are difficult to be applied to many datasets if some specific information is required. Therefore, we consider the ranking task for summarsing the model performance on the asking clarification question task and present the results of BM25 and Doc2Query + BM25. Note that, the BM25-based techniques are considered with their competitive performance in addressing the clarification question ranking task (Aliannejadi et al., 2021). We also compare some additional ranking

[13] https://scikit-learn.org/

[14] https://huggingface.co/

techniques, such as the PL2 (Amati and Van Rijsbergen, 2002), DPH (Amati et al., 2008) and another recent dense retriever (i.e., ColBERT (Khattab and Zaharia, 2020)). However, the inclusion of such approaches is not useful while comparing the use of different datasets. Therefore, we only present the results of the above two approaches in Table 5. As for the implementation, we leverage PyTerrier[15] (Macdonald and Tonellotto, 2020), a recently developed Python framework for conducting information retrieval experiments.

B.3 User Satisfaction with CQs

In this experiment, we explored the task of determining user satisfaction with CQs by utilizing a variety of models from both traditional machine learning and language models on the ACQs datasets. To conduct this experiment, we employed the same models that we previously used for the Clarification Need Prediction task. By using the same models for both tasks, we aim to examine how well these models perform in predicting user satisfaction with CQs and how their performance compares to their performance in predicting the need for clarification. This will allow us to understand the strengths and limitations of these models in predicting user satisfaction and make informed decisions on which models to use in future applications. Only two datasets (i.e., MIMICS (Zamani et al., 2020) and MIMICS-Duo (Tavakoli et al., 2022)) out of 12 datasets provide the user satisfaction information. In both MIMICS and MIMICS-Dou, each clarification question is given a label to indicate how a user is satisfied with the clarification question. For MIMICS the labels are Good, Fair, or Bad. A good clarifying question is accurate, fluent, and grammatically correct. A fair clarifying question may not meet all of these criteria but is still acceptable. Otherwise, it is considered bad. While in MIMICS-Dou, users' satisfaction with clarification questions is assessed on a 5-level scale that is Very Bad, Bad, Fair, Good, and Very Good. Thus, we formulate user satisfaction with CQs task as a supervised classification in our experiments.

[15] https://github.com/terrier-org/pyterrier

Table 7: The performance of all methods on clarification need prediction on MIMICS and MIMICS-Duo. The best models are in **bold**.

Model	**MIMICS**			**MIMICS-Duo**		
	Precision	Recall	F1	Precision	Recall	F1
RandomForest	0.3540	**0.3806**	**0.3717**	0.2860	0.5000	0.3638
DecisionTree	0.2125	0.2520	0.2028	0.5329	0.5095	0.4305
SVM	0.2858	0.3024	0.2772	0.5281	0.5088	0.4333
MultinomialNB	0.2924	0.3186	0.2876	0.5185	0.5178	0.5166
LogisticRegression	0.2749	0.2878	0.2816	**0.7862**	0.5010	0.3660
BART	0.5083	0.3344	0.3657	0.5869	0.5503	0.5194
XLNet	0.1385	0.2500	0.1782	0.286	0.5	0.3638
XLM	0.0119	0.2500	0.0227	0.286	0.5	0.3638
Albert	0.2920	0.2877	0.2855	0.286	0.5	0.3638
distilBERT	0.3391	0.3305	0.3322	0.5941	0.594	0.5941
BERT	**0.3804**	0.3249	0.3344	0.6349	**0.625**	**0.6255**
	MIMICS			**MIMICS-Duo**		
	MAE	MSE	R^2	MAE	MSE	R^2
RandomForest	2.4404	7.969	-0.0012	2.8502	11.206	*-0.0079*
DecisionTree	2.6374	10.0143	-0.2581	3.052	14.2306	-0.2799
SVR	2.4447	8.1852	-0.0283	2.7801	14.6398	-0.3167
MultinomialNB	3.3364	16.7424	-1.1034	2.7971	18.942	-0.7037
LogisticRegression	3.4084	17.9488	-1.2549	2.7971	18.942	-0.7037
BART	2.3903	8.5296	-0.0716	**2.7233**	**10.3239**	**0.0714**
XLNet	2.4582	8.1836	-0.0281	2.7971	18.942	-0.7037
XLM	2.6214	9.9151	-0.2456	2.7971	18.942	-0.7037
Albert	2.4339	8.0300	-0.0088	2.7971	18.942	-0.7037
distilBERT	**2.3325**	**7.8685**	**0.0115**	2.7744	11.0613	0.0051
BERT	2.4562	8.1277	-0.0211	2.8801	11.2268	-0.0098

Table 8: The performance of all methods on user satisfaction prediction with CQs on MIMICS and MIMICS-Duo. The best models are in **bold**.

Model	**MIMICS**			**MIMICS-Duo**		
	Precision	Recall	F1	Precision	Recall	F1
RandomForest	0.7522	0.5172	0.3686	0.1256	0.25	0.1672
DecisionTree	0.5648	0.5168	0.4050	0.2218	0.2311	0.2163
SVM	0.736	0.5947	0.5212	0.2379	0.2498	0.2157
MultinomialNB	0.8255	0.7842	0.7758	**0.4407**	0.2787	0.2336
LogisticRegression	0.7522	0.5172	0.3686	0.3762	0.2542	0.1761
BART	0.9385	0.931	0.9302	0.1256	0.25	0.1672
XLNet	0.9219	0.9217	0.9217	0.1256	0.25	0.1672
XLM	0.9348	0.9309	0.9303	0.1256	0.25	0.1672
Albert	0.9385	0.931	0.9302	0.1256	0.25	0.1672
distilBERT	**0.9453**	**0.9397**	**0.939**	0.2766	**0.2803**	**0.2777**
BERT	0.9385	0.931	0.9302	0.2851	0.264	0.2056

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Section 6

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.