MIMICS: A Large-Scale Data Collection for Search Clarification

Hamed Zamani
Microsoft
hazamani@microsoft.com

Gord Lueck
Microsoft
gordonl@microsoft.com

Everest Chen
Facebook Inc.
everestch@fb.com

Rodolfo Quispe
Microsoft
edquispe@microsoft.com

Flint Luu
Microsoft
flintl@exchange.microsoft.com

Nick Craswell
Microsoft
nickcr@microsoft.com

ABSTRACT

Search clarification has recently attracted much attention due to its applications in search engines. It has also been recognized as a major component in conversational information seeking systems. Despite its importance, the research community still feels the lack of a large-scale data for studying different aspects of search clarification. In this paper, we introduce MIMICS, a collection of search clarification datasets for real web search queries sampled from the Bing query logs. Each clarification in MIMICS is generated by a Bing production algorithm and consists of a clarifying question and up to five candidate answers. MIMICS contains three datasets: (1) MIMICS-Click includes over 400k unique queries, their associated clarification panes, and the corresponding aggregated user interaction signals (i.e., clicks). (2) MIMICS-ClickExplore is an exploration data that includes aggregated user interaction signals for over 60k unique queries, each with multiple clarification panes. (3) MIMICS-Manual includes over 2k unique real search queries. Each query-clarification pair in this dataset has been manually labeled by at least three trained annotators. It contains graded quality labels for the clarifying question, the candidate answer set, and the landing result page for each candidate answer.

MIMICS is publicly available for research purposes,[1] thus enables researchers to study a number of tasks related to search clarification, including clarification generation and selection, user engagement prediction for clarification, click models for clarification, and analyzing user interactions with search clarification.

1 INTRODUCTION

Search clarification has recently been recognized as a useful feature for improving user experience in search engines, especially for ambiguous and faceted queries [29]. In addition, it has been identified as a necessary step towards developing mixed-initiative conversational search systems [22, 28]. The reason is that limited bandwidth interfaces used in many conversational systems, such as speech-only and small-screen devices, make it difficult or even impossible for users to go through multiple documents in case of ambiguous or faceted queries. This has recently motivated researchers and practitioners to investigate possible approaches to *clarify* user information needs by asking a question [1, 29].

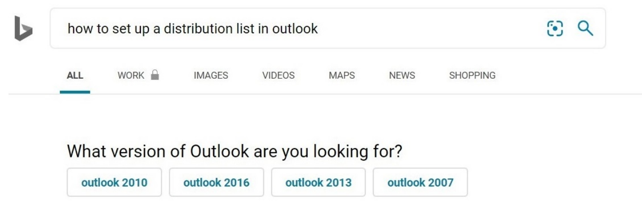


Figure 1: An example of clarification pane in Bing.

Despite the recent progress in search clarification, e.g., [1, 13, 29, 30], the community still feels the lack of a large-scale dataset for search clarification, which is necessary for speeding up the research progress in this domain. To address this issue, we introduce MIMICS,[2] a data collection consisting of multiple datasets for search clarification. Each clarification in MIMICS consists of a clarifying question and up to five candidate answers. Figure 1 shows the interface used for clarification in Bing for constructing this data. The first dataset, called MIMICS-Click, includes over 400k unique search queries sampled from the Bing's query logs, each associated with a single clarification pane. The dataset also includes aggregated user interaction signals, such as the overall user engagement level and conditional clickthrough rate on individual candidate answers. The second dataset, called MIMICS-ClickExplore, contains over 64k queries, each with multiple clarification panes which are the result of multiple exploration and online randomization experiments. This dataset also includes the aggregated user interaction signals. The third dataset, on the other hand, is manually labeled by trained annotators. This dataset, which is called MIMICS-Manual, includes graded quality labels for clarifying question, candidate answer set, and the landing result page for each individual answer.

The datasets created as part of MIMICS can be used for training and evaluating a variety of tasks related to search clarification, including generating/selecting clarifying questions and candidate answers, re-ranking candidate answers for clarification, click models for search clarification, user engagement prediction for search clarification, and analyzing user interactions with search clarification. This paper also suggests some evaluation methodologies and metrics for these tasks.

[1] MIMICS is available at https://github.com/microsoft/MIMICS.

[2] MIMICS stands for the Microsoft's Mixed-Initiative Conversation Search Data.

This work was done while Everest Chen was affiliated with Microsoft.

Microsoft, July 2020, Technical Report

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Clarification has been explored in a number of applications, such as speech recognition [26], dialogue systems [4, 12, 20], and community question answering [5, 23, 24]. Recently, it attracted much attention in the information retrieval literature [1, 13, 22, 29, 30]. For instance, Kiesel et al. [15] investigated the impact of voice query clarification on user satisfaction. Their study showed that users like to be prompted for clarification. Simple form of clarification, such as entity disambiguation, has been explored by Coden et al. [8]. They basically ask a "did you mean A or B?" question to resolve entity ambiguity. Even earlier, Allan [2] organized the HARD Track at TREC 2004 which involved clarification from participants. In more detail, the participants could submit a form containing some human-generated clarifying questions in addition to their submission run. Recently, Aliannejadi et al. [1] proposed studying clarification in the context of conversational information seeking systems. This was later highlighted as an important aspect of conversational search in the Dagstuhl Seminar on Conversational Search [3]. More recently, Zamani et al. [29] introduced clarification in the context of web search and proposed models for generating clarifying questions and candidate answers for open-domain search queries. In a follow-up study, Zamani et al. [30] analyzed user interactions with clarification panes in Bing and provided insights into user behaviors and click bias in the context of search clarification. Moreover, Hashemi et al. [13] proposed a representation learning model for utilizing user responses to clarification in conversational search systems.

Despite the recent progress reviewed above, there is no large-scale publicly available resource for search clarification. To the best of our knowledge, Qulac[3] [1] is the only public dataset that focuses on search clarification. However, it only contains 200 unique queries borrowed from the TREC Web Track 2009-2012. Therefore, it is not sufficient for training a large number of machine learning models with millions of parameters. In addition, it was constructed through crowdsourcing. Therefore, the clarifications are human generated and user responses to clarifications in real scenarios may differ from the ones in Qulac. There also exist a number of community question answering data and product catalogs with clarifications (e.g., see [24]), however, they are fundamentally different from search clarification. Therefore, this paper provides a unique resource in terms of realisticness, size, diversity, clarification types, user interaction signals, and coverage.

It is worth noting that a number of datasets related to conversational search has recently been created and released. They include CCPE-M [21], CoQA [25], QuAC [7], MISC [27], and the Conversation Assistance Track data created in TREC 2019 [11]. Although these datasets do not particularly focus on clarification, there might be some connections between them and MIMICS that can be used in future research. In addition, the public query logs, such as the one released by AOL [19], can be used together with MIMICS for further investigations. This also holds for the datasets related to query suggestion and query auto-completion.

3 DATA COLLECTION

Bing has recently added a clarification pane to its result pages for some ambiguous and faceted queries. It is located right below the search bar and above the result list. Each clarification pane includes a clarifying question and up to five candidate answers. The user interface for this feature is shown in Figure 1. The clarifying questions and candidate answers have been generated using a number of internal algorithms and machine learning models. They are mainly generated based on users' past interactions with the search engine (e.g., query reformulation and click), content analysis, and a taxonomy of entity types and relations. For more information on generating clarification panes, we refer the reader to [29] that introduces three rule-based and machine learning models for the task. All the datasets presented in this paper follow the same properties and only demonstrate the queries from the en-US market.

In the following subsections, we explain how we created and pre-processed each dataset introduced in the paper. In summary, MIMICS consists of two datasets (MIMICS-Click and MIMICS-ClickExplore) based on user interactions (i.e., clicks) in Bing and one dataset (MIMICS-Manual) based on manual annotations of clarification panes by multiple trained annotators.

3.1 MIMICS-Click

We sub-sampled the queries submitted to Bing in September 2019. We only kept the queries for which a clarification pane was rendered in the search engine result page (SERP). We made efforts in our data sampling to cover a diverse set of query and clarification types in the dataset, therefore, the engagement levels released in the paper by no mean represent the overall clickthrough rates in Bing. For privacy reasons, we followed k-anonymity by only including the queries that have been submitted by at least 40 users in the past year. In addition, the clarification panes were solely generated based on the submitted queries, therefore they do not include session and personalized information. We performed additional filtering steps to preserve the privacy of users using proprietary algorithms. Sensitive and inappropriate contents have automatically been removed from the dataset. To reduce the noise in the click data, we removed the query-clarification pairs with less than 10 impressions. In other words, all the query-clarification pairs released in the dataset have been presented at least 10 times to the Bing users in the mentioned time period (i.e., one month).

This resulted in 414,362 unique queries, each associated with exactly one clarification pane. Out of which 71,188 of clarifications have received positive clickthrough rates. The statistics of this dataset is presented in Table 1.

The dataset is released in a tab-separated format (TSV). Each data point in MIMICS-Click is a query-clarification pair, its impression level (low, medium, or high), its engagement level (between 0 and 10), and the conditional click probability for each individual candidate answer. The engagement level 0 means there was no click on the clarification pane. We used a equal-depth method to divide all the positive clickthrough rates into ten bins (from 1 to 10). The description of each column in the dataset is presented in Table 2.

3.2 MIMICS-ClickExplore

Although MIMICS-Click is a invaluable resource for learning to generate clarification and related research problems, it does not allow researchers to study some tasks, such as studying click bias in user interactions with clarification. Therefore, to foster research in these

[3] https://github.com/aliannejadi/qulac

Table 1: Statistics of the datasets constructed as part of MIMICS.

	MIMICS-Click	MIMICS-ClickExplore	MIMICS-Manual
# unique queries	414,362	64,007	2464
# query-clarification pairs	414,362	168,921	2832
# clarifications per query	1 ± 0	2.64 ± 1.11	1.15 ± 0.36
min & max clarifications per query	1 & 1	2 & 89	1 & 3
# candidate answers	2.81 ± 1.06	3.47 ± 1.20	3.06 ± 1.05
min & max # candidate answers	2 & 5	2 & 5	2 & 5
# query-clarification pairs with positive engagement	71,188	89,441	N/A
# query-clarification pairs with low/medium/high impressions	264,908 / 105,879 / 43,575	52,071 / 60,907 / 55,943	N/A

Table 2: The data format in MIMICS-Click and MIMICS-ClickExplore.

Column(s)	Type	Description
query	string	The query text
question	string	The clarifying question
option_1, $\cdots$, option_5	string	The candidate answers from left to right. If there is less than five candidate answers, the rest would be empty strings.
impression_level	string	A string associated with the impression level of the corresponding query-clarification pair. Its value is either 'low', 'medium', or 'high'.
engagement_level	integer	An integer from 0 to 10 showing the level of total engagement received by the users in terms of clickthrough rate.
option_cctr_1, $\cdots$, option_cctr_5	real	The conditional click probability on each candidate answer. They must sum to 1, unless the total_ctr is zero. In that case, they all are zero.

interesting and practical tasks, we created MIMICS-ClickExplore using some exploration and randomization experiments in September 2019. In more detail, we used the top m clarifications generated by our algorithms and presented them to different sets of users (similar to A/B testing). The user interactions with multiple clarification panes for the same query at the same time period enable comparison of these clarification panes. The difference between these clarification panes can be in the clarifying question, the candidate answer set, the order of candidate answers, or a combination of them.

We performed the same filtering approach to address privacy concerns as the one discussed above for MIMICS-Click. Again, we only kept the query-clarification pairs with a minimum impression of 10. The resulted dataset contains 64,007 unique queries and 168,921 query-clarification pairs. Out of which, 89,441 query-clarification pairs received positive engagements.

The format of this dataset is the same as MIMICS-Click (see Table 2). Note that the sampling strategies for MIMICS-Click and MIMICS-ClickExplore are different which resulted in significantly more query-clarification pairs with low impressions in MIMICS-Click.

3.3 MIMICS-Manual

Although click provides a strong implicit feedback signal for estimating the quality of models in online services, including search clarification, it does not necessarily reflect all quality aspects. In addition, it can be biased for many reasons. Therefore, a comprehensive study of clarification must include evaluation based on manual human annotations. This has motivated us to create and release MIMICS-Manual based on manual judgements performed by trained annotators.

Therefore, we randomly sampled queries from the query logs to collect manual annotations for a set of realistic user queries. The queries satisfy all the privacy concerns reviewed in Section 3.1. We further used the same algorithm to generate one or more clarification pairs for each query. Each query-clarification pair was assigned to at least three annotators. The annotators have been trained to judge clarification panes by attending online meetings, reading comprehensive guidelines, and practicing. In the following, we describe each step in the designed Human Intelligence Task (HIT) for annotating a query-clarification pair. This guideline has been previously used in [29, 30].

3.3.1 Step I: SERP Review. Similar to Aliannejadi et al. [1], we first asked the annotators to skim and review a few pages of the search results returned by Bing. Since search engines try to diversify the result lists, this would enable the annotators to better understand the scope of the topic and different potential intents behind the submitted query. When completed, the users can move to the next step.

3.3.2 Step II: Annotating the Clarifying Question Quality. In this step, the annotators were asked to assess the quality of the given clarifying question independent of the candidate answers. Therefore, the annotation interface does not show the candidate answers to the annotators at this stage. Each clarifying question is given a label 2 (Good), 1 (Fair), or 0 (Bad). The annotators were given detailed definitions, guidelines, and examples for each of the labels. In summary, the guideline indicates that a Good clarifying question should accurately address and clarify different intents of the query.

Table 3: The data format in MIMICS-Manual. All the labels in this dataset are either 2 (Good), 1 (Fair), or 0 (Bad).

Column(s)	Type	Description
query	string	The query text
question	string	The clarifying question
option_1, $\cdots$, option_5	string	The candidate answers from left to right. If there is less than five candidate answers, the rest would be empty strings.
question_label	integer	The label associated with the clarifying question independent of the candidate answers.
options_overall_label	integer	The overall label given to the candidate answer set.
option_label_1, $\cdots$, option_label_5	integer	The label assigned to each individual candidate answer based on the quality of the landing search result page.

Table 4: The statistics of the common clarifying question templates in MIMICS. We only present the templates with at least 100 occurrence in MIMICS-Click and MIMICS-ClickExplore individually. Note that there is no label associated with the first template in MIMICS-Manual.

ID	Clarifying question template	MIMICS-Click		MIMICS-ClickExplore		MIMICS-Manual	
		Freq.	Engagement	Freq.	Engagement	Freq.	Question Quality
T1	select one to refine your search	395134	0.9285	156870	2.8631	2490	N/A
T2	what (do you want|would you like) to know about (.+)?	7136	0.5783	5624	2.9070	158	1.9367
T3	(which|what) (.+) do you mean?	7483	0.6123	1905	2.6714	76	2.000
T4	(what|which) (.+) are you looking for?	3436	1.7252	2055	5.1990	22	1.6818
T5	what (do you want|would you like) to do with (.+)?	689	1.9637	1833	3.4043	60	2.000
T6	who are you shopping for?	101	1.9604	350	4.3800	7	1.5714
T7	what are you trying to do?	188	3.3777	116	5.8793	3	1.0

Table 5: The average and standard deviation of user engagement levels with respect to different query-clarification impressions.

Impression level	MIMICS-Click	MIMICS-ClickExplore
Low	0.9061 ± 2.5227	3.1712 ± 4.2735
Medium	0.9746 ± 2.1249	3.1247 ± 3.3622
High	0.9356 ± 1.6159	2.4119 ± 2.4559

It should be fluent and grammatically correct. If a question fails in satisfying any of these factors but still is an acceptable clarifying question, it should be given a Fair label. Otherwise, a Bad label should be assigned to the question. Note that if a question contains sensitive or inappropriate content, it would have been flagged by the annotators and removed from the dataset. Note that in case of having a generic template instead of clarifying questions (i.e., "select one to refine your search"), we do not ask the annotators to provide a question quality labels.

3.3.3 Step III: Annotating the Candidate Answer Set Quality. Once the clarifying question is annotated, the candidate answers would appear on the HIT interface. In this step, the annotators were asked to judge the overall quality of the candidate answer set. In summary, the annotation guideline indicates that the candidate answer set should be evaluated based on its usefulness for clarification, comprehensiveness, coverage, understandability, grammar, diversity, and importance order. A clear definition of each of these constraints has been mentioned in the guideline. Note that the annotators have reviewed multiple pages of the result list in Step I and have been expected to know different possible intents of the query. Again, the labels are either 2 (Good), 1 (Fair), or 0 (Bad), and the candidate answers with sensitive or inappropriate contents have been removed from the dataset. If a candidate answer set satisfies all the aforementioned constraints, it should be given a Good label. While, the Fair label should be given to an acceptable candidate answer set that does not satisfy at least one of the constraints. Otherwise, the Bad label should be chosen. Note that since all the defined properties are difficult to satisfy with up to 5 candidate answers, the label Good is rarely chosen for a candidate answer set.

3.3.4 Step IV: Annotating the Landing SERP Quality for Each Individual Candidate Answer. Zamani et al. [29] recently performed a number of user studies related to search clarification. In their interviews, the participants mentioned that the quality of the secondary result page (after clicking on a candidate answer) perceived the usefulness of the clarification pane. Based on this observation, we asked the annotators to evaluate the quality of the secondary result page (or the landing result page) for the individual candidate answers one by one. Therefore, the annotators could click on each individual answer and observe the secondary result page in Bing. Since a SERP may contain multiple direct answers, entity cards, query suggestion, etc. in addition to the list of webpages, adopting ranking metrics based on document relevance, such as mean reciprocal rank (MRR) or normalized discounted cumulative gain (NDCG) [14], is not desired to evaluate the overall SERP quality. Therefore, we again asked the annotators to assign a label 2 (Good), 1 (Fair), or 0 (Bad) to each landing SERP. A label Good should be chosen, if the correct answer to all possible information needs behind the selected candidate answer can be easily found in

Table 6: The average and standard deviation of engagement levels and manual annotation labels per query length.

Query	MIMICS-Click		MIMICS-ClickExplore		MIMICS-Manual			
length	Freq.	Engagement	Freq.	Engagement	Freq.	Question quality	Answer set quality	Landing page quality
1	52213	0.5158 ± 1.6546	26926	1.9508 ± 2.7098	1028	1.7347 ± 0.4415	1.0418 ± 0.3075	1.9750 ± 0.1251
2	160161	0.7926 ± 2.1548	70621	2.7965 ± 3.3536	942	1.4694 ± 0.4991	1.0085 ± 0.3827	1.9178 ± 0.2881
3	120821	1.0152 ± 2.4573	46070	3.1677 ± 3.5811	555	1.4667 ± 0.4989	0.9333 ± 0.4463	1.8021 ± 0.4816
4	51503	1.2196 ± 2.6980	16798	3.5397 ± 3.7492	199	1.3333 ± 0.4714	0.9698 ± 0.5103	1.8313 ± 0.41986
5	19893	1.4473 ± 2.9078	5755	4.0188 ± 3.8921	75	1.3846 ± 0.4865	1.0267 ± 0.5157	1.7847 ± 0.5291
6	6299	1.5785 ± 3.0318	1806	4.1877 ± 3.9642	15	1.0 ± 0.0	0.8 ± 0.5416	1.7 ± 0.4800
7	2424	1.6634 ± 3.0815	621	4.6715 ± 3.9861	13	1.0 ± 0.0	0.7692 ± 0.4213	1.7692 ± 0.5756
8	823	1.7618 ± 3.1575	264	4.2008 ± 3.9082	3	N/A	1.0 ± 0.0	1.8333 ± 0.2357
9	184	1.9620 ± 3.2959	52	4.1731 ± 3.8467	1	N/A	0.0 ± 0.0	2.0 ± 0.0
10+	41	2.0732 ± 3.4244	8	4.8750 ± 3.4799	1	N/A	1.0 ± 0.0	2.0 ± 0.0

a prominent location in the page (e.g., an answer box on top of the SERP or the top three retrieved webpages). If the result page is still useful and contain relevant information, but finding the answer is not easy or is not on top of the SERP, the Fair label should be selected. Otherwise, the landing SERP should be considered as Bad.

3.3.5 A Summary of the Collected Data. Each HIT was assigned to at least three annotators. For each labeling task, we used majority voting to aggregate the annotation. In case of disagreements, the HIT was assigned to more annotators. The overall Fleiss' kappa inter-annotator agreement is 63.23%, which is considered as good.

Our annotations resulted in over 2.4k unique queries and over 2.8k query-clarification pairs. The statistics of the dataset is reported in Table 1. The data has been released in a tab-separated file format (TSV). The description of each column in the data is provided in Table 3.

4 DATA ANALYSIS

In this section, we provide a comprehensive analysis of the created datasets.

4.1 Question Template Analysis

Zamani et al. [29] showed that most search clarifications can be resolved using a small number of question templates. In our first set of analysis, we study the question templates in MIMICS and their corresponding statistics. We only focus on the templates with a minimum frequency of 100 in both MIMICS-Click and MIMICS-ClickExplore. We compute the average engagement level per clarifying question template for MIMICS-Click and MIMICS-ClickExplore. In addition, we compute the average question quality label per template for MIMICS-Manual that has manual annotations. Note that engagement levels are in the $[0, 10]$ interval, while the manual annotation labels are in $[0, 2]$. The results are reported in Table 4. The first general template is excluded in our manual annotations. According to the results, the last four templates (T4 - T7) have led to higher engagements compared to T1, T2, and T3 in both MIMICS-Click and MIMICS-ClickExplore. They are also generally less frequent in the dataset and more specific. In general, the exploration dataset has higher average engagements compared to MIMICS-Click. The reason is that the number of query-clarification pairs with zero engagements in MIMICS-Click are higher than those in MIMICS-ClickExplore (see Table 1).

4.2 Analyzing Engagement Based on Clarification Impression

As mentioned in Section 3, MIMICS-Click and MIMICS-ClickExplore contain a three-level impression label per query-clarification pair. The impression level is computed based on the number of times the given query-clarification pair has been presented to users. The impression level should have a correlation with the query frequency. We compute the average and standard deviation of engagements per impression level whose results are reported in Table 5. According to the results, there is a negligible difference between the average engagements across impression levels. Given the engagements range (i.e., $[0, 10]$), the query-clarification pairs with high impressions in MIMICS-ClickExplore have led to slightly lower average engagements.

4.3 Analysis Based on Query Length

In our third analysis, we study user engagements and manual quality labels with respect to query length. To this aim, we compute the query length by simply splitting the query using whitespace characters as delimiters. The results are reported in Table 6. According to the results on MIMICS-Click and MIMICS-ClickExplore, the average engagement increases as the queries get longer. By looking at the data one can realize that longer queries are often natural language questions, while short queries are keyword queries. Surprisingly, this is inconsistent with the manual annotations suggesting that single word queries have higher question quality, answer set quality, and also landing page quality (excluding the rare queries with less than 10 frequency in the dataset). This observation suggests that user engagement with clarification is not necessarily aligned with the clarification quality. The behavior of users who submit longer queries may differ from those who search with keyword queries.

4.4 Analysis Based on the Number of Candidate Answers

As pointed out earlier, the number of candidate answers in the data varies between two and five. To demonstrate the impact of

Table 7: The average and standard deviation of engagement levels and manual annotation labels per number of candidate answers.

# answers	MIMICS-Click		MIMICS-ClickExplore		MIMICS-Manual			
	Freq.	Engagement	Freq.	Engagement	Freq.	Question quality	Answer set quality	Landing page quality
2	226697	0.9047 ± 2.3160	50474	2.8430 ± 3.3921	1083	1.3164 ± 0.4651	0.9751 ± 0.3775	1.8915 ± 0.3665
3	91840	0.9904 ± 2.4175	38619	3.0592 ± 3.5111	892	1.7513 ± 0.4323	0.9507 ± 0.2954	1.9129 ± 0.3101
4	42752	0.9276 ± 2.3505	29678	2.9157 ± 3.4395	453	1.6292 ± 0.4830	1.0088 ± 0.3816	1.9073 ± 0.2862
5	53073	0.9099 ± 2.3323	50150	2.8354 ± 3.4236	404	1.4741 ± 0.4993	1.1733 ± 0.5401	1.9168 ± 0.2832

the number of candidate answers, we report the average and standard deviation of engagement levels and manual quality labels per number of candidate answers in Table 7. According to the results, there is a small difference between average engagements in both MIMICS-Click and MIMICS-ClickExplore datasets. The clarifications with three candidate answers have led to a slightly higher engagement than the rest. It is again in contrary to the manual quality labels; the clarifications with three candidate answers have obtained the lowest answer set quality label. On the other hand, the question quality of clarifications with three candidate answers is higher than the others. This highlights that the question quality may play a key role in increasing user engagements.

4.5 Analyzing Click Entropy Distribution on Candidate Answers

MIMICS-Click and MIMICS-ClickExplore both contain conditional click probability on each individual answer, i.e., the probability of clicking on each candidate answer assuming that the user interacts with the clarification pane. The entropy of this probabilistic distribution demonstrates how clicks are distributed across candidate answers. The entropy range depends on the number of candidate answers, therefore, we normalized the entropy values by the maximum entropy per the candidate answer size. The distribution for MIMICS-Click and MIMICS-ClickExplore are reported in Figures 2 and 3, respectively. Note that for the sake of visualization, these plots do not include clarifications with no click (i.e., the engagement level zero) and those with zero entropy. According to the plots, the number of peaks in the entropy distribution is aligned with the number of candidate answers. The entropy values where the histogram peaks suggest that in many cases there is a uniform-like distribution for m out of n candidate answers (for all values of m). Comparing the plots in Figure 2 with those in Figure 3 shoes that this finding is consistent across datasets.

5 INTRODUCING RESEARCH PROBLEMS RELATED TO SEARCH CLARIFICATION

MIMICS enables researchers to study a number of research problems. In this section, we introduce these tasks and provide high-level suggestions for evaluating the tasks using MIMICS.

5.1 Clarification Generation

Clarification generation (including both clarifying question and candidate answers) is a core task in search clarification. Generating clarification from a passage-level text has been studied in the context of community question answering posts [24]. It has lately attracted much attention in information seeking systems, such as search engines (similar to this study) [29] and recommender systems [32]. Previous work has pointed out the lack of large-scale training data for generating search clarification [1, 29]. MIMICS, especially the click data, provides an excellent resource for training clarification generation models.

Evaluating clarification generation models, on the other hand, is difficult. One can use MIMICS for evaluating the generated clarification models using metrics such as BLEU [18] and ROUGE [17]. However, we strongly discourage this evaluation methodologies, as they poorly correlate with user satisfaction and clarification quality. Here is our recommendation for evaluating clarification generation models:

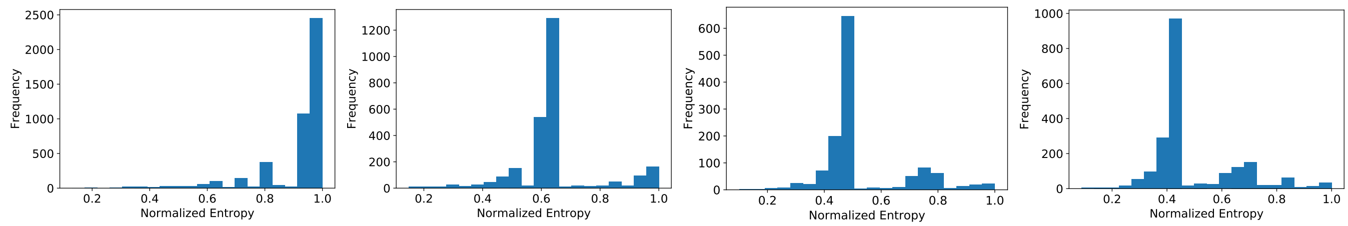
- In case of access to production systems with real users, conducting online experiments (e.g., A/B tests) would be a reliable evaluation methodology and the models can be compared using user engagement measures, such as clickthrough rate.
- Manual annotation of the generated clarifications based on carefully-defined criteria would be an alternative for clarification generation evaluation. Previously, Zamani et al. [29] used this evaluation methodologies. Researchers may adopt the annotation guideline presented in Section 3.3 for designing their crowdsourcing HITs.

5.2 Clarification Selection

Since automatic offline evaluation of clarification generation models is difficult, clarification selection (or clarification re-ranking) can be considered as an auxiliary task to evaluate the quality of learned representations for clarification. In addition, as pointed out by Aliannejadi et al. [1], information seeking systems can adopt a two stage process for asking clarification, i.e., generating multiple clarifications and selecting one. Selecting clarification has been previously studied in [1, 13, 30].

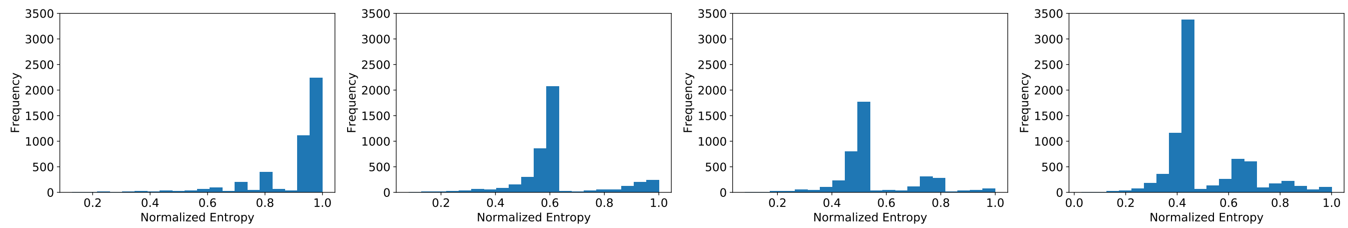
Researchers can benefit from MIMICS for both training and evaluating clarification selection models. In more detail, MIMICS-ClickExplore contains multiple clarifications per query and can be directly used for evaluating clarification selection (or re-ranking) models. The other two datasets can be also used by drawing some negative samples that can be obtained either randomly or using a baseline model.

Ranking metrics, such as NDCG, can be used to evaluate clarification selection models. In addition, since only one clarification is often shown to the users, the average engagement of the selected clarification can be also chosen as an evaluation metric. Refer to [30] for more information.



(a) # candidate answers = 2 (b) # candidate answers = 3 (c) # candidate answers = 4 (d) # candidate answers = 5

Figure 2: The distribution of normalized entropy for the conditional clickthrough rates on candidate answers for the MIMICS-Click dataset. For the sake of clarity and visualization, we exclude the clarification with no click and those with zero entropy.



(a) # candidate answers = 2 (b) # candidate answers = 3 (c) # candidate answers = 4 (d) # candidate answers = 5

Figure 3: The distribution of normalized entropy for the conditional clickthrough rates on candidate answers for the MIMICS-ClickExplore dataset. For the sake of clarity and visualization, we exclude the clarification with no click and those with zero entropy.

5.3 User Engagement Prediction for Clarification

A major task in search clarification is deciding whether to ask clarification, especially in search systems with limited-bandwidth interfaces. This problem can be cast to query performance prediction [6, 10]. In other words, clarification can be asked when the predicted performance for the given query is below a threshold. An alternative to query performance prediction for this task would be user engagement prediction. In more detail, if users enjoy interacting with clarification and find it useful, the system can decide to ask the clarification. Predicting user engagement has been previously studied in various contexts, such as social media and web applications [16, 31], however, user engagement prediction for clarification is fundamentally different. MIMICS-Click and MIMICS-ClickExplore contain engagement levels in the $[0, 10]$ interval. Therefore, they can be directly used for predicting user engagements.

For evaluating user engagements prediction models for clarification, we recommend computing correlation between the predicted engagements and the actual observed engagement released in the datasets. Correlation has been also used for evaluating query performance prediction models [6]. Since we only release engagement levels, we suggest using both linear (e.g., Pearson's ρ) and rank-based (e.g., Kendall's τ) correlation metrics.

In addition, mean square error or mean absolute error can be used for evaluating user engagement prediction methods.

5.4 Re-ranking Candidate Answers

Previous work has shown that the order of candidate answers in clarification matters [30]. MIMICS enables researchers to study the task of re-ranking candidate answers for a given pair of query and clarifying question. Experiments on both click data (MIMICS-Click and MIMICS-ClickExplore) and manual annotations would provide complementary evaluation for the task.

For evaluating the candidate answers re-ranking task, the manual annotations per individual answers based on their landing SERP quality can be used as graded relevance judgement. NDCG would be adopted as the evaluation metric. For evaluation using the click data, researchers should be careful about presentation bias in the data. Refer to [30] for more detail. In summary, the candidate answers with higher ranks and longer text are more likely to attract clicks. This point should be considered prior to using the MIMICS-Click and MIMICS-ClickExplore for re-ranking candidate answers. Once this issue is addressed, the conditional click probabilities can be mapped to ordinal relevance labels and typical ranking metrics can be adopted for evaluation. One can also use cross-entropy between the predicted probability distribution for candidate answers and the actual conditional click distribution. The impression level can be also considered in the metric to compute a gain per query-clarification pair with respect to their impression. In more detail, the clarifications that are presented more often should be assigned higher weights.

5.5 Click Models for Clarification

Related to the re-ranking candidate answers task, it is important to design user models for their click behavior while interacting with clarification panes. Zamani et al. [30] showed that the existing click models that have primarily been designed for web search do not perform as expected for search clarification. The reason is that the assumptions made in the web search click models do not hold for search clarification. The MIMICS-ClickExplore dataset contains many clarification pairs for a given query whose only differences are in the order of candidate answers. This allows researchers to train and evaluate click models for search clarification using MIMICS-ClickExplore. The evaluation methodology used in [30] is suggested for evaluating the task. In summary, it is based on predicting the click probability of swapping adjacent candidate answers. This approach has originally been used for evaluating click models in web search by Craswell et al. [9]. The cross-entropy would be an appropriate metric in this evaluation setup.

5.6 Analyzing User Behavior in Search Clarification

Although this paper provides several analyses based on search clarification quality in terms of both manual judgements and engagement levels, future work can benefit from MIMICS-Click and MIMICS-ClickExplore to conduct more in depth analysis of user behaviors while interacting with search clarification in the context of web search.

6 CONCLUSIONS

In this paper, we introduced MIMICS, a data collection for studying search clarification, which is an interesting and emerging task in the context of web search and conversational search. MIMICS was constructed based on the queries and interactions of real users, collected from the search logs of a major commercial web search engine. MIMICS consists of three datasets: (1) MIMICS-Click includes over 400k unique queries with the associated clarification panes. (2) MIMICS-ClickExplore is an exploration data and contains multiple clarification panes per query. It includes over 60k unique queries. (3) MIMICS-Manual is a smaller dataset with manual annotations for clarifying questions, candidate answer sets, and the landing result page after clicking on individual candidate answers. We publicly released these datasets for research purposes.

We also conducted a comprehensive analysis of the user interactions and manual annotations in our datasets and shed light on different aspects of search clarification. We finally introduced a number of key research problems for which researchers can benefit from MIMICS.

In the future, we intend to report benchmark results for a number of standard baselines for each individual task introduced in the paper. We will release the results to improve reproducibility and comparison. There exist a number of limitations in the released datasets. For instance, they only focus on the en-US market and do not contain personalized and session-level information. These limitations can be resolved in the future.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR '19*. 475–484.

[2] James Allan. 2004. HARD Track Overview in TREC 2004: High Accuracy Retrieval from Documents. In *TREC '04*.

[3] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search - A Report from Dagstuhl Seminar 19461. *arXiv preprint arXiv:2005.08658* (2020).

[4] Marco De Boni and Suresh Manandhar. 2003. An Analysis of Clarification Dialogue for Question Answering. In *NAACL '03*. 48–55.

[5] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly?: Analyzing Clarification Questions in CQA. In *CHIIR '17*. 345–348.

[6] D. Carmel and E. Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval* (1st ed.). Morgan and Claypool Publishers.

[7] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP '18*. 2174–2184.

[8] Anni Coden, Daniel Gruhl, Neal Lewis, and Pablo N. Mendes. 2015. Did you mean A or B? Supporting Clarification Dialog for Entity Disambiguation. In *SumPre '15*.

[9] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *WSDM '08*. 87–94.

[10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2002. Predicting Query Performance. 299–306.

[11] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. TREC CAsT 2019: The Conversational Assistance Track Overview. In *TREC '19*.

[12] Marco De Boni and Suresh Manandhar. 2005. Implementing Clarification Dialogues in Open Domain Question Answering. *Nat. Lang. Eng.* 11, 4 (2005), 343–361.

[13] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. In *SIGIR '20*.

[14] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[15] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *SIGIR '18*. 1257–1260.

[16] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. 2014. *Measuring User Engagement.* Morgan & Claypool Publishers.

[17] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics. In *NAACL '03*. 71–78.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL '02*. 311–318.

[19] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search. In *InfoScale '06*.

[20] Luis Quintano and Irene Pimenta Rodrigues. 2008. Question/Answering Clarification Dialogues. In *MICAI '08*. 155–164.

[21] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *SIGDIAL '19*.

[22] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *CHIIR '17*. 117–126.

[23] Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *ACL '18*. 2737–2746.

[24] Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *NAACL '19*.

[25] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL* 7 (2019), 249–266.

[26] Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2014. Towards Natural Clarification Questions in Dialogue Systems. In *AISB '14*, Vol. 20.

[27] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *CAIR '17*.

[28] Hamed Zamani and Nick Craswell. 2020. Macaw: An Extensible Conversational Information Seeking Platform. In *SIGIR '20*.

[29] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *WWW '20*. 418–428.

[30] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and Learning from User Interactions with Search Clarification. In *SIGIR '20*.

[31] Hamed Zamani, Pooya Moradi, and Azadeh Shakery. 2015. Adaptive User Engagement Evaluation via Multi-Task Learning. In *SIGIR '15*.

[32] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *CIKM '18*. 177–186.