1 ANALYSIS

To intuitively present our ConvSearch dataset, we conduct a series of statistical analyses in this section. Table 1 shows the basic statistics of ConvSearch dataset, where *# total turns* denotes the number of turns in the original dialogue data, while *# merged turns* denotes the number of turns after we merge multiple consecutive responses of user or agent into a single turn. Next, we will present more detailed analysis of the ConvSearch dataset.

1.1 Statistical Characteristics of Dialogue Content

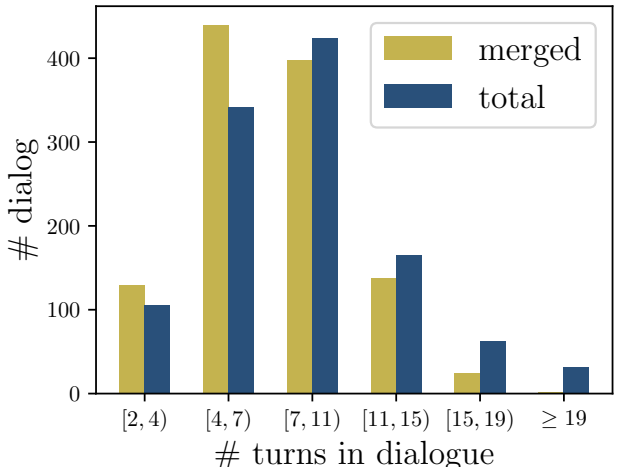


Figure 1: The distribution of the number of dialogue turns in ConvSearch dataset

Figure 1 shows the distribution of the number of dialogue turns in our dataset. We can find that the majority number of total (68%) and merged (74%) turns in a dialogue range from 4 to 10. Figure 2 depicts the topic distribution of dialogue in ConvSearch. Except for the topic *politics* with a slightly fewer proportion, the number of dialogues under different dialogue topic is relatively even.

Figure 3a and Figure 3b show the distribution on the text content length[1] of merged turns initiated by user and agent respectively. The significant difference in the average content length between users and agents deserves our attention: the mean value is 15.5 for users, while 440 for agents. This phenomenon is reasonable. Because users' request expressed via keyboards is usually short, although it is still longer than that in web search context[2]. While agents are more likely to extract or generate longer answers based on the relevant information they have sought for.

As for the format of the agent responses, 6.27% of the turns contain link in content, while image has been adopted in 10.00% of the turns.

1.2 Statistics of Dialogue Annotation Results by Users and Agents

Table 2 shows the distribution of dialogue-level and turn-level annotation results assessed by users and agents. More than eighty percent of users rate *satisfaction_user* at a high level (greater than 2), which reflects agents are qualified with the ability of understanding and responding users' information need by communicating with them and searching relevant information in replace of them. In most dialogue situations, agents own solid ability to understand the information needs of users, and then acquire information to satisfy user information

[1]Here, we ignore the merged turns which only contain information in image format.

Table 1: Basic statistics of ConvSearch dataset

Statistics	Number
# Dialogues	1,131
Average total turns per dialogue	8.30
Average merged turns per dialogue	6.89
Average agent queries per dialogue	6.01
Average agent queries with clicks per dialogue	3.68
Average words per agent query	10.37
Average agent landing pages per dialogue	8.78

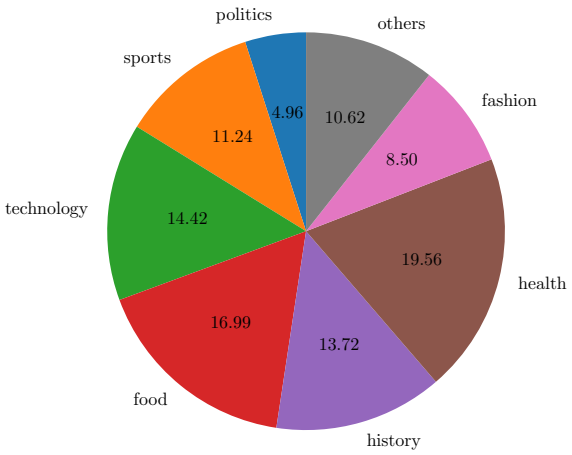


Figure 2: The distribution of the dialogue topic in ConvSearch dataset

Table 2: The distribution of dialogue-level (top) and turn-level (bottom) annotation results assessed by users and agents. The boldface highlights the most frequent annotation grade. The value represents the percentage of the category.

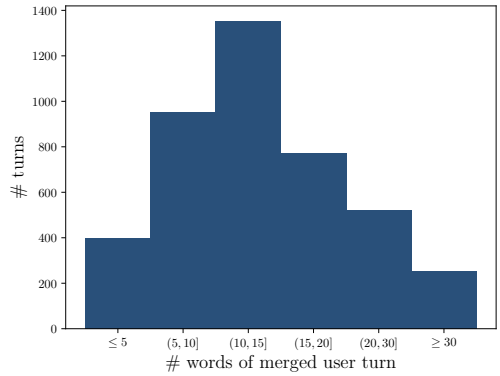
Annotation item	0	1	2	3	4
satisfaction_user	0.44	2.12	8.67	31.33	**57.43**
applicability_user	0.66	4.81	12.45	38.87	**43.21**
effort_user	**32.04**	30.27	22.48	11.86	3.36
preference_user[2]	11.06	29.65	**42.83**	8.23	8.23
difficulty_agent	23.21	21.35	**26.66**	20.81	7.97
understand_agent	0.00	0.27	4.16	12.40	**83.17**
satisfaction-turn	0.75	2.04	8.68	21.74	**66.79**
understand-turn	0.55	1.09	4.50	15.28	**78.58**
clarity-turn	0.39	1.01	4.47	11.54	**82.58**
difficulty-turn	**40.91**	18.31	18.22	15.09	7.46

needs. Turn-level user satisfaction is even higher than dialogue-level. As for *effort_user*, over half of users feel that they have paid a low price to get their information need satisfied by conversational search. This is the phenomenon we enjoy seeing: high satisfaction at low cost in terms of effort devoted. From the agents' viewpoint, the perceived difficulty of users' information need is more uniformly distributed in Table 2. It shows that our dataset covers various situations: conversation tasks with different levels of difficulty can be found in our dataset.

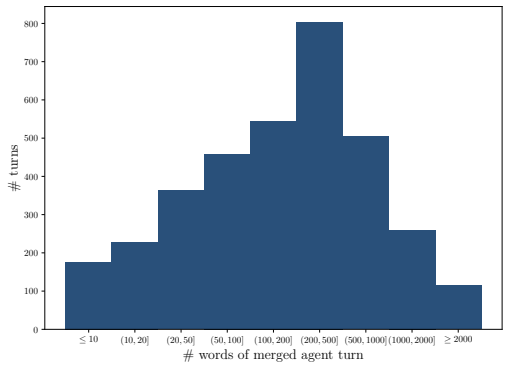
1.3 Statistics of Agent Search Behaviors

Query and result related behaviors constitute the two major components of agent search behaviors. As for SERPs, Figure 4 shows some distributions of it. Figure 4a shows that more than two third of the SERPs have no more than one document clicked by agents. This means that in most cases agents can have a clear picture of whether the results under the SERP will assist them in their information seeking tasks and thus make a definitive decision between clicking or leaving. Figure 4b shows the click-through rate distribution under different ranking positions on SERPs. The results show a significant drop in the click-through rate of the document as the ranking position increases. This is partly due to the diminishing relevance of the results as the ranking position increases, but also consistent with the findings of the position bias discovered by the previous work on click model [1].

As for landing pages, Table 3 shows the distribution on the number of landing pages with different levels of *helpfulness-pagelog* under a merged turn. The results show that for more than half of the turns, the agent clicks on at least one very useful (*helpfulness-pagelog* equals to



(a) merged turns initiated by user



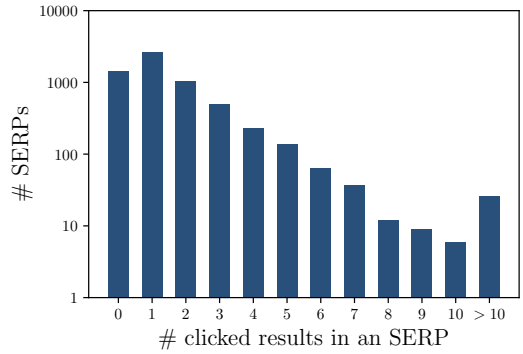
(b) merged turns initiated by agent

Figure 3: The distribution on the text content length of merged turns initiated by user and agent in ConvSearch dataset

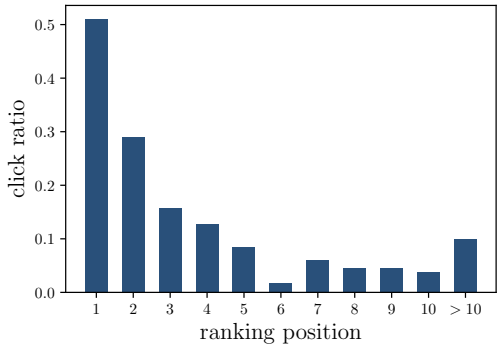
Table 3: The distribution on the number of landing pages with different levels of *helpfulness-pagelog*. For example, the number 231 in the third row of the fifth column means that there exist 231 merged turns that are generated by the agent after viewing 2 landing pages whose *helpfulness-pagelog* label is 3.

# pagelogs	grade-0	grade-1	grade-2	grade-3	grade-4
0	2,978	2,808	2,691	2,332	1,474
1	170	276	341	544	1,353
2	71	114	157	231	300
3	31	43	66	113	94
4	37	48	45	60	54
5	14	17	16	25	27
6	13	13	9	12	12
7	7	8	4	14	6
≥ 8	34	28	26	24	35

4) document before in responses to user. Only less than 17% of the responses are completed before the agent clicks on any landing pages. We suspect that these responses mainly belong to the *chit-chat* and *clarify* classes.



(a) Distribution on the number of documents clicked under SERPs



(b) Click-through rate distribution under different ranking positions

Figure 4: The click distribution under Search Engine Result Pages (SERPs)

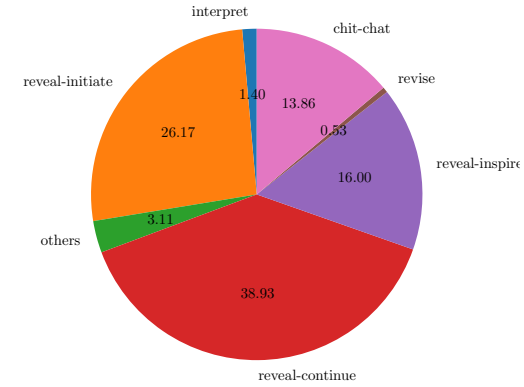
1.4 Statistics of User Intent and Agent Action Annotation Results

Since some turns in terms of user intent and agent action are annotated by means of the dual primary and secondary labels, here we only analyze the primary label for simplicity. The Kappa values of the primary labels in the annotation of user intent (7 classes) and agent action (7 classes) by three assessors are 0.8013 and 0.4774 respectively. *answer-single-fact*, *answer-multi-fact* and *answer-opinion* are the most confusing categories for assessors when annotating agent actions. This significant difference indicates that the labeling of agent action is more difficult than that of user intent. Figure 5 shows the distributions of primary labels for user intent and agent action annotation. It is not difficult to find that *reveal-continue* and *answer-multi-fact* occupy the main categories of user intent and agent action respectively.

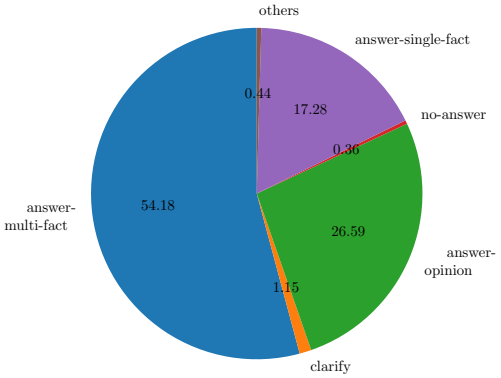
REFERENCES

[1] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.

[2] Zhijing Wu, Xiaohui Xie, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. A study of user image search behavior based on log analysis. In *China Conference on Information Retrieval*. Springer, 69–80.



(a) User intent



(b) Agent action

Figure 5: The categorization distributions of primary labels for user intent and agent action annotation