Categorizing User Messages in Text-based Informational Conversations

ANONYMOUS AUTHOR(S)

Online text-based informational conversation (such as the chats between customers and representatives) is an important way of addressing people's information needs. We have concluded a categorization system for user messages in informational conversations after a thorough examination of a real-world customer service log. The responses included both human and AI chatbot's outputs. We categorize user messages into five categories (including 15 specific types) related to three high-level intentions. For example, users can ask new questions, restate previous questions, and provide supplement details to describe their information needs. They can also communicate with the other side (such as clarification and giving feedback) to ensure they understand the received information. Also, a substantial set of messages does not describe needs or discuss information but helps maintain the conversation (such as chit-chats and nudge messages). Two annotators independently classified the same set of 1,478 user messages from 300 conversations and reached a moderate consistency (Cohen's Kappa 0.59). We summarize and report the characteristics of different message types and compare their usage in sessions with only human, AI, or both representatives. Our results show that different message types vary significantly in usage frequency, length, and text similarities with other messages in a session. Also, the frequency of using different message types in our dataset seems consistent over sessions with different types of representatives. But we also observed some significant differences in a few specific message types across the sessions with different representatives, suggesting it is necessary to examine further the influence of human, AI, and mixed representatives on user messages and informational conversation.

Additional Keywords and Phrases: chatbot, conversational agent, query reformulation, customer service

1 INTRODUCTION

Finding and acquiring information is an ever-lasting need of human society. The primary portal of information access today is web search engines. Web search engines provide easy-to-use and flexible search interface and access to billions of entries on the Internet. However, web search engines cannot address all information needs. An increasing popular alternative service today is conversational agents (such as Siri, Google, Alexa, and Cortana). They provide information to users in a conversation model, allowing people to communicate in a natural way. In addition to the free conversational agents provided to the public, many commercial companies also provide human-chatbot hybrid customer services to address customers' needs, as many of the information provided by customer services may be internal and not accessible on the public web.

An important area of study in web search engines is query reformulation—how searchers modify queries in a session and why. Previous studies have concluded many common patterns, e.g., a query may add content to or remove words from a past query, may substitute a part of the past query with new words, may reorder words sequence, etc. Some patterns may also connect to search intentions, such as to search more specific or more generic topics, or to switch to different subtopics under the same theme etc. These studies provide fundamental basis for understanding search engine user behavior and designing related techniques, e.g., automatic query

suggestion and completion, and online search evaluation methods (as some query reformulation patterns may also indicate search quality). Similarly, previous studies also examined patterns of changing user requests in conversational dialog systems, but they focused on not only content changes but also phonetic differences, e.g., a user's request may repeat but overstate a previous utterance. Many studies had also classified user requests and system responses into dialog acts that indicate different intentions.

We focus on studying online text-based informational conversation. It refers to the online communications that users chat with a conversational agent in text to address their information needs. We do not consider voice communications in this study, though many conversational agents provide multi-modal inputs and outputs. Text-based informational conversation is an important addition to search engines for people to acquire information, e.g., online customer services. The possible advantages of chat-based informational conversation are that they can be more natural and interactive to users. Also, conversational search may provide more efficient information access particularly when it is difficult for the users to retrieve the requested information from web search engines (e.g., the users may not be able to figure out effective search keywords but they can describe their information needs to the agent in detail).

Particularly, we are interested in what types of messages users send in informational conversations, and the underlying intentions behind these messages. We follow previous studies of search engine query reformulation and dialog system acts and conclude a classification scheme of user messages based on thorough examination of a real-world customer service log. Our classification scheme included five categories of messages (including 15 specific types) linked to three higher-level intentions: describing information, understanding information, and maintaining conversation. Two annotators independently classified the same set of 1,478 user messages from 300 conversations and came to a moderate consistency with Cohen's Kappa = 0.59. Also, the customer service system employed a hybrid model involving both human representatives and chatbot outputs. Thus, our dataset also included sessions with only human, AI, or mixed representatives—this makes it possible to also examine the differences of messages and message types in sessions with different representatives.

We are particularly interested in the following research questions:

- **RQ1—**What types of messages do users send in text-based information conversations? What are the possible intentions behind sending these message types?
- **RQ2—**How do different types of messages vary in their characteristics and usage frequency?
- **RQ3—**Does the use of different message types vary in sessions with different representatives?

The rest of this article introduces our categorization scheme, data annotation process, and results.

2 RELATED WORK

2.1 Search Engine Query Reformulation

Query reformulation refers to the activity of formulating a query that is different from an existing one, where the focus is the difference of the two queries. Previous work characterized many patterns of reformulation [12, 13, 14, 17, 18, 20, 22]. Some are characterized from the lexical aspect, for example: adding words, removing words, replacing words by synonyms, spelling correction, stemming, case change, and using acronyms. Some are concerned with syntactic differences, for example: punctuation, reordering words, and using search operators. Some patterns may imply users' intents, for example: specification, generalization, and subtopic change. These

patterns are not exclusive of each other. For example, one can reformulate a more specific query (specification) either by adding words or replacing words with more specific ones.

In addition to categorize query reformulation patterns, previous work also examined the choices of words in query reformulation and interactive relevance feedback. For example, Spink et al. [24] studied five sources of query terms in mediated online searching. Among the sources they examined, question statement is similar to task description in our study, and we also consider relevance feedback as a source for new terms. In addition, the content of search results is also an important source of knowledge for query reformulation [25]. Yue et al. [21] examined possible sources of query words in collaborative search. Some of them may also be applied to other types of searches, including users' past queries and viewed search results. Another source we examined is query suggestions displayed on the SERP. Kelly et al. [19] compared term and query suggestions, where users reported that query suggestions provide ideas for manually formulating queries.

Also, previous studies also developed technical solutions for contextual search and query suggestion based on these word changes. Guan et al. [16] separately considered added, retained, and removed words in relevance feedback; Dang et al. [15] generated synthesized query suggestions by considering similar patterns; Awadallah et al. predicted voice query reformulation patterns [11].

2.2 Dialogue Acts in Conversational Dialog Systems

Dialogue acts [9, 10] are fine-grained classification systems for user-system communications in conversational dialog systems. Previous studies developed different dialogue acts categories with different granularity, and used the acts to manage conversations, generate responses, model users, and evaluate systems.

Core and Allen [1] proposed a task-independent annotation schema Dialogue Act Marking in Several Layers (DAMSL) in 1997. The annotation scheme is fine-grained with 220 tags divided into four categories depending on their roles in the conversation and characteristics. Stolcke et al. [3] introduced an approach for modeling dialogue acts in conversational speech, which could detect and predict dialogue acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialogue act sequence. They focus on recognizing 42 major dialogue act types from Jurafsky et. al.'s work [2], such as Statements and Opinions, Questions, Backchannels, Turn Exits, Answers and Agreements, etc. In addition to users' dialogue acts, systems' dialogue acts could also be classified. For example, Walker et al. [4] described a dialogue act tagging scheme that classified each system utterance into categories such as request-info, present, offer, acknowledge, status-report, explicit-confirm, implicit-confirm, instruct, apologize, open/close, etc. In contrast to user acts in a dialogue, system acts usually do not need to be recognized because they can be defined while designing the system. Dialogue acts also have many applications in commercial products. For example, Walker et al. [23] used dialog acts and patterns to evaluate spoken dialog systems, and Jiang et al. [8] defined several dialogue acts in intelligent voice assistant and used their transition patterns to predict conversation quality.

Approaches of defining and analyzing dialog acts have also been applied to examine human discourse of other types. For example, Ferschke et al. [5] proposed an annotation schema for the discourse analysis of Wikipedia Talk pages to improve the article. They applied the annotation schema to a corpus of 100 Talk pages from the Simple English Wikipedia and performed automatic dialog act classification on Wikipedia discussions. Oraby et al. [6] developed a taxonomy of fine-grained dialogue acts frequently observed in customer service dialogue on Twitter. The taxonomy contained Greeting, Statement, Request, Question, Answer, Social Act. They modeled conversation flow and predicted the dialogue act of a given turn in real-time by using a sequential

SVM-HMM model. Braslavski et al. [7] studied the clarification questions asked by CQA users in two different domains, analyzed their behavior, and the types of clarification questions asked. Our work also follows previous studies of dialogue acts but apply similar methods to examine online text-based conversations.

3 ASKER MESSAGE CATEGORY IN ONLINE INFORMATIONAL CONVERSATIONS

3.1 Informational Conversation Sessions

We examine online text-based conversations between two parties. Particularly, we focus on the conversations where the primary goal is to address one party's information need, and the other party provides information. We call such conversations *informational conversations*. Informational conversation is an important method to address people's information needs, especially in commercial services such as online customer support. Other types of conversations also exist—for example, a conversation can also be transactional (e.g., chatting with a colleague to make an appointment), discussional (e.g., debating with a friend about presidential candidates), or entertaining (e.g., conversations that are just for fun)—but we do not discuss them here.

We further define the two parties of an informational conversation:

- ***Asker***—the party who hopes to address an information need from a conversation.
- ***Answerer***—the party who provides information to the asker during the conversation.

We only consider the case where the conversation includes one asker and one answerer. We note that the key difference between an asker and an answerer is not whether they ask questions or not, but their roles for addressing information needs. As we will discuss in the following sections, an answerer may also ask questions during a conversation to help an asker describe the information need.

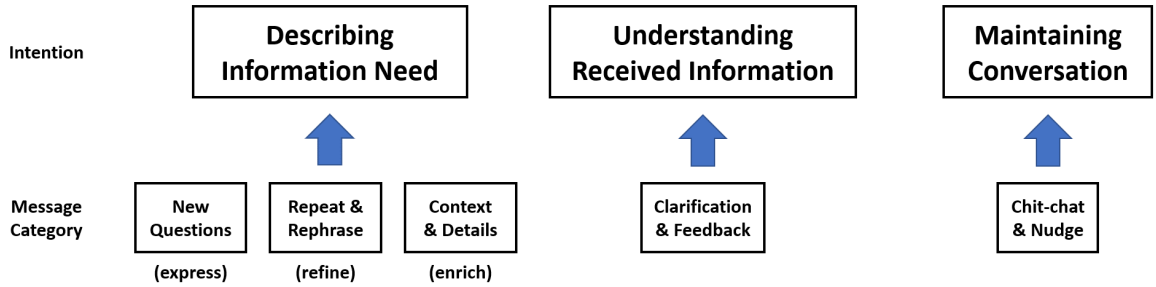

Figure 1: Higher-level asker message categories and related asker intentions during a conversation.

3.2 Asker Message Categories and Intentions

We categorize asker messages in informational conversations into five higher-level categories and 15 specific types and summarize three possible intentions. Figure 1 illustrates how the five higher-level message categories relate to three possible intentions:

- ***Describing information need***—A substantial number of asker messages aims to describe information needs, which is not necessarily a one-shot process. On one hand, an asker may have several related but different needs. On the other hand, a single message may not convey a need well. We concluded three message categories for describing information needs, including those for asking new questions, refining previous questions, and providing contexts and supplementary details.

- ***Understanding received information***—Some of the asker messages' intention is to understand the information received from the answerer. This includes messages for clarifying the received responses and providing feedback to the answerer's responses.
- ***Maintaining conversation***—Some of the asker messages do not directly address information needs but help maintain an active conversation, e.g., chit-chat messages such as greetings, nudge messages that check if the answerer is available.

We summarize this classification system based on two annotators' examination of a real-world dataset and previous studies of search engine query reformulation types and conversational dialog system acts. The dataset includes informational conversations between customers (askers) and online representatives (answerers) from a health insurance provider company in China. The representatives had both human workers and AI chatbots—we decided not to categorize answerer responses as those sent by human representatives and AI chatbots are very different. All messages are in Chinese (Mandarin). We report examples translated into English in the paper and enclose the original Chinese messages in the appendix for reference.

3.2.1 New Questions

An asker message may describe a new question that the same information need has not been asked before in the conversation session. We further divide such messages into three types depending on whether and how these messages relate to past messages in the same session.

- ***Question (Q)***—a question that is not a follow up of any previous messages.
- ***Follow-up Question, Self (FQS)***—a question following up on a past message sent by the asker.
- ***Follow-up Question, Answerer (FQA)***—a question following up on a past response from the answerer.

Table 1 and 2 show example messages that were classified as follow-up questions. We use shaded cells for asker messages in all the following examples. The message in Table 1 is classified as FQS because it is related to the previous asker message regarding health insurance coverage with fewer than ¥ 10,000 annual expenses. The message in Table 2 is classified as FQA because it is following up the answerer's response mentioning claiming reimbursement afterwards.

Table 1: An example message classified as **Follow-up Question, Self (FQS)**.

Q	So, it means that I won't get compensated if I don't spend over ¥ 10,000 a year, right?
	Correct.
FQS	*I should ask social security or other health insurance for compensation if less than ¥ 10,000, right?*
	Yes, correct.

Table 2: An example message classified as **Follow-up Question, Answerer (FQA)**.

Q	If I am hospitalized, do I need to pay for the expenses myself first?
	Hello, this insurance is a reimbursement insurance. You need to pay for the expenses first yourself and claim reimbursement afterwards from the insurance company with your supporting documents.
FQA	*How long is afterwards?*
	Sorry, I can't understand what you meant.

3.2.2 Repeat & Rephrase

An asker message may also restate a question where the same information need has been asked before in the session. We further divide such messages into four types depending on the difference between the message and previous ones stating the same information need:

- ***Repeat***—restating a previous question without any change.
- ***Rephrase***—restating a previous question with only wording differences.
- ***Rephrase, Add***—restating a previous question with more content.
- ***Rephrase, Delete***—restating a previous question with less content.

Table 3, 4, and 5 show some example messages that were classified as rephrased questions in our dataset. The key difference between rephrased questions and FQS (a question following up a previous asker question) is whether the same information need has been asked before (despite any content difference). As the example messages show, a rephrased question states the same information with some previous asker messages, even though some content may have been added or removed. In contrast, the example FQS message in Table 1 is related to a previous Q message, but it conveys a different question. The reason for repeating and rephrasing questions is mostly because the answerer was not able to provide effective responses.

Table 3: An example message classified as **Rephrase**.

Q	Will my insurance fee be returned?
	Our insurance has a high compensation rate. You can pay as low as ¥ 100 to be compensated up to ¥ 6 million. You can check the insurance fee for different ages by clicking on micro app—health insurance—estimate my first-year insurance fee.
Rephrase	*I was asking if I can get my insurance fee back.*

Table 4: An example message classified as **Rephrase, Add**.

Q	Will it be covered if I am hospitalized?
	Our insurance is a compensatory heal insurance. The covered expenses mainly include hospitalization expenses, specialized clinics expenses, surgery expenses, ER expenses before and after hospitalization. Our coverage is not limited to hospitalization expenses.
Rephrase, Add	*I do not have an insurance yet. I am hospitalized now. Will it be covered?*

Table 5: An example message classified as **Rephrase, Delete**.

Q	My child is 3. Should I check the no social security option?
	Please wait a second while I am answering your question.
	Hello! Both people with and without social security are eligible for our insurance, but the fees are different.
Rephrase, Delete	*So, child has no social security, right?*

3.2.3 Contexts and Details

An asker's message may not ask a question but provide contexts and details to enrich information needs. Such messages can be either self-initiated or elicited by the answerer. We summarized four types of such messages:

- ***Background (BG)***—asker's self-initiated messages providing background information *before* a question.
- ***Supplement (SUP)***—asker's self-initiated messages providing supplement information *after* a question.
- ***Answer (ANS)***—asker's messages responding to questions from the answerer.
- ***Correction***—correcting typos or incorrect details in a previous message.

Table 6 and 7 show examples that were classified as Background and Supplement messages in our dataset. We define Background and Supplement messages as those that are not questions and only describe contexts or details. In contrast, askers may also include more contexts and details while rephrasing a question, but such rephrased messages are stating questions. We suspect that an important reason for sending Background and Supplement messages is that users may not prefer drafting long messages in online text chatting, especially on mobile devices. In such a case, it is natural to split a long question into separate messages, where some of the messages may only provide contexts and details. Table 7 also shows a message classified as Answer. This message also provides context to the problem but is elicited by the answerer.

Correction messages take a very small fraction of our dataset (0.9%). They include both messages that only rectify the incorrect part of a previous message and those restating a corrected question. Theoretically, we can further conclude an individual type "Rephrase, Correction" for the latter case. But here we simply count all these messages into one type since they are very rare in our dataset.

Table 6: An example message classified as **Background (BG)**.

Background	*I was hospitalized for fracture last year*
	You are eligible for the insurance if you have been hospitalized in the past two years for the following reasons: A) labor; B) acute respiratory diseases; C) acute gastroenteritis or appendicitis; D) gallstones that did not relapse in two years; E) benign gallbladder polyps; F) accidental hospitalization recovered in 5 days without sequelae or loss of any organ.
Q	Am I not eligible?

Table 7: Example messages classified as **Supplement (SUP)** and **Answer (ANS)**.

Q	Will the insurance rate increase every year?
	As one gets older, the risk of having an accident or disease increases too, and the insurance rate also increases. But our insurance aims to be inclusive. Even for people over 60 years old, they only need to pay a little more than ¥ 1,000 a year (i.e., roughly ¥ 100 monthly rate) to have an insurance that can compensate up to ¥ 6 million. It is a highly cost-effective product that everyone can afford.
Supplement	*I am 49 years old.*
	Please wait a second while I am answering your question.
	May I ask if you have social security? [rose emoji]
Answer	*Yes, I have.*
	Hello, the rate for 49 years old with social security is ¥ 858 a year.

3.2.4 Clarification & Feedback

An asker's message may not describe an information need but to communicate with the answerer to better understand the responses. We have concluded two types of messages for this purpose:

- ***Clarify (CLR)***—clarifying the meaning of the answerer's responses.
- ***Feedback***—providing (negative) feedback and requests to the answerer.

Table 8 shows an example message classified as Clarify (CLR) in our dataset. A Clarify message may also be a question, which makes it look similar to a follow-up question related to the Answerer (FQA). However, the key difference between them is that a Clarify message does not convey new information need, but a follow-up question does express one by our definition.

Table 9 shows an example message classified as Feedback in our dataset. The message provides feedback to the previous responses and requests for a simpler answer. Feedback messages take only a small fraction of

all the messages in our dataset (0.9%). Most of them expressed negative feelings towards answerer responses. The purpose of the Feedback messages is to notify answerers and urge them to improve response quality.

3.2.5 Chit-chat & Nudge

We have also identified many chit-chat messages and a very small number of nudge messages in the dataset. Chit-chat includes individual messages such as greeting and goodbye messages. We define nudge messages as those trying to notify the answerer and urge them to respond (e.g., is anyone there, are you still there). These messages do not describe information need, but they are useful for keeping the conversation natural and active.

Table 8: An example asker message classified as **Clarify (CLR)**.

Q	How soon I will be covered after buying this insurance?
	There is a 30-day waiting period starting the midnight of the next day you have purchased our insurance.
	The waiting period is also called the observation period, which refers to the 30-day period after purchasing the insurance. During this period, the customer's medical expenses caused by diseases will not be covered. The purpose of having a waiting period is to guarantee the benefits of healthy customers and to avoid people buying this insurance while they have already been diagnosed with certain diseases such that we can ensure the benefits of the majority.
Clarify	*So, I can only be covered after 30 days*
	Correct.

Table 9: An example asker message classified as **Feedback**.

Q	Okay, effective starting the midnight of the next day of purchasing, may I explain this item as that I can get covered for serious diseases starting the next day after I purchased this insurance?
	Hello, customers are not covered for diseases within 30 days after purchasing this insurance.
	But the waiting period does not include accidents [happy emoji]
Feedback	*Sorry, I can't understand, can you make it simpler?*
	Please wait a second while I am answering your question.
	To make it simpler: after successfully purchasing our insurance, you have no waiting period for accident compensation, but there is a 30-day waiting period if you are hospitalized for diseases, which is the same as the observation period you mentioned [happy emoji]
	Does not make more sense to you?

4 DATA AND ANNOTATION

4.1 System

Our dataset comes from a company's online text-based customer support log. The company is a primary health insurance provider in China. Thus, all the conversations are related to health insurance. The company provides online customer support through a popular mobile messaging app in China. Figure 2 shows a screenshot of the messaging interface for a customer. The right-aligned messages (green boxes) are from a customer, and the left-aligned messages (white boxes) are from a service representative. On the bottom is a text input box for the customer. Customers can communicate with the representatives in the same way they send a regular message in this app. The company also provides customer services on other platforms, but our dataset only included communications via this messaging app.

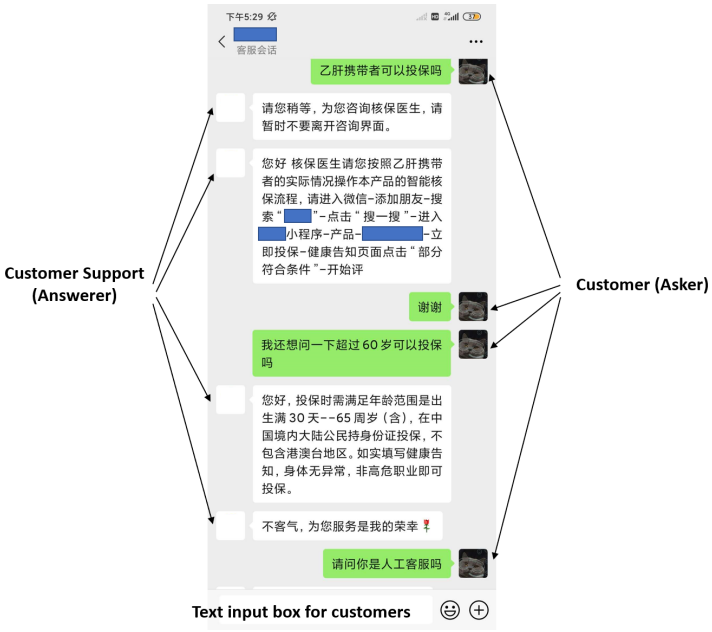

Figure 2: A screenshot of the messaging interface for customers (askers). We hide the names of the insurance company and products (blue boxes) and the avatar of the customer support due to anonymization requirements.

The company's online customer service employs a hybrid model, including a group of human representatives and an AI chatbot. For an incoming customer message, the system first computes a chatbot message response. If the chatbot's confidence level (an internal measure indicating the quality of the generated message) is below a certain threshold, the system will try to switch to a human representative. If all human representatives are busy, the system will let the customer wait. If the waiting time exceeds a threshold, the system will send an automatic response, "Please wait for a second while I am answering your question," to notify the customer. We also count this automatic message into the conversation as customers did receive them. If all human workers are offline (such as during nighttime), the system will respond with AI chatbot messages even if they have low confidence scores. The system does not indicate customers if they are talking to human representatives or an AI chatbot. However, we observed that some customers inquired and requested to switch to human services.

4.2 Dataset

We created a dataset of informational conversations based on the company's customer service conversation log. The log includes conversations between October 2017 and January 2018. We further divide a conversation session into three types based on the types of representatives involved:

- **AI-only**—all the responses are from an AI chatbot.
- **Human-only**—all the responses are from a human representative.
- **Hybrid**—the session includes responses from both AI chatbot and human representative.

We randomly sampled 100 sessions for each type. The dataset includes 300 conversations, involving 1,478 customer messages and 1,936 representative responses. We examined all the conversations manually to make sure they are informational conversations. We have excluded sessions where customers had sent multiple

consecutive messages without receiving any responses (0.6% of the sessions in the log belongs to this type). We further define a *round* of a conversation as the period from one asker message (inclusive) to the next asker message (exclusive). Thus, each round in our dataset includes one and only one asker message but may have one or multiple answerer responses.

Note that our selection of AI-only, Human-only, and Hybrid sessions is quasi-experimental. Particularly, a conversation may end up being AI-only just because the chatbot has high confidence scores for all customer messages. Thus, we suspect the complexity and difficulty of customer questions in the three types of sessions may vary. One should be cautious when reading our results comparing the three session types, because the differences may not entirely come from the different types of representatives involved in the conversations.

4.3 Annotation Procedure and Consistency

Two of the authors independently annotated the dataset to categorize asker messages into different types. All the messages are using Chinese Mandarin, and both annotators are also native Chinese Mandarin speakers to make sure they can correctly understand the messages.

We (including both annotators and another author) first discussed and came out with an initial classification scheme based on a small sample of the data (including 100 messages). The initial scheme also borrowed ideas from previous studies of search engine query reformulation and dialog system acts. The initial scheme had included 11 types and did not involve Background, Supplement, Correction, and Feedback. The two annotators discussed cases that could not categorize into the initial scheme during the annotation process and had gradually enriched the scheme to the form we introduced in Section 3.

The two annotators' results have a moderate consistency—the overall Cohen's Kappa on the whole dataset (including three types of sessions) is 0.59. The agreements are higher in human-only sessions (0.67) but lower in hybrid ones (0.55). Further, the two annotators discussed the disagreed messages and came to agreements. Table 10 reports the consistency between each annotator's results and the final agreed types after discussion.

For 11 out of the 15 message types (excluding Q, CH, Nudge, and Feedback), the two annotators had also identified their related messages/responses in the session. For example, the related message of an FQS is the asker's previous message that the FQS followed up, and the related message of a CLR is the response that the asker was trying to clarify. The two annotators also had high agreements on the identified related messages. Among those messages where both annotators agreed on the message type, they also agreed on 83.6% of the related messages.

Table 10: Cohen's Kappa among the two annotators' results and the final results after discussion.

	ALL	AI-only	Human-only	Hybrid
Cohen's Kappa: Annotator 1 vs. 2	0.59	0.58	0.67	0.55
Cohen's Kappa: Final vs. Annotator 1	0.74	0.71	0.82	0.73
Cohen's Kappa: Final vs. Annotator 2	0.80	0.83	0.84	0.77
% agreed related messages	83.6%	81.2%	85.7%	84.2%

5 MESSAGE TYPES AND CHARACTERISTICS

5.1 Overall Statistics

Table 11 reports overall statistics about sessions, rounds, and messages in our dataset. The results suggest that the three types of sessions in our dataset are very different from many aspects.

First, askers had used much more rounds to finish conversations in the Hybrid sessions (7.26 on average, in contrast to 4.42 for AI-only and 3.10 for Human-only sessions). These differences consequently made Hybrid sessions differ greatly with the other two types in the number of messages and responses at a session level, although askers had received significantly more responses during a round in Human-only sessions (1.73) than in Hybrid ones (1.32). Note that every AI session round included consistently one response because the chatbot is designed to always respond only one message for each request. We suspect the high number of rounds in a session may indicate that customers had low conversation quality in Hybrid sessions, as previous studies had also identified long search sessions and dialog sessions as negative signals for search/conversation quality.

Second, the messages and responses from AI and Human sessions also differ significantly in length (by the number of Chinese characters), though we found neither of them had any significant difference with messages in Hybrid sessions. The length of responses also differs greatly between AI-only and human-only sessions. This suggests that AI and human representatives are providing very different responses, and asker messages may also be different in these two session types (13.08 vs. 15.61 characters).

To conclude, many statistics in Table 11 showed that the conversations and messages in the three session types have lots of differences in our dataset. Such differences may come from that the influence of the different types of representatives in these sessions, our selection criterion when building this dataset, or both. Yet further investigation is needed to better understand such differences.

Table 11: Mean and standard error of various statistics for sessions, rounds, and messages in different sessions. We test significant differences using a one-way ANOVA with the Tukey HSD post hoc test.

User Message Category	ALL	AI-only	Human-only	Hybrid	$P < 0.05$ Differences
# messages & responses / session	11.38 (0.54)	8.84 (0.76)	8.47 (0.42)	16.83 (1.19)	Hybrid > AI, Human
# asker messages / session	4.93 (0.24)	4.42 (0.38)	3.10 (0.15)	7.26 (0.51)	Hybrid > AI > Human
# AI responses / session	2.37 (0.19)	4.42 (0.38)	-	2.68 (0.26)	AI > Hybrid
# human responses / session	4.09 (0.28)	-	5.37 (0.28)	6.89 (0.59)	Hybrid > Human
# rounds / session	4.93 (0.24)	4.42 (0.38)	3.10 (0.15)	7.26 (0.51)	Hybrid > AI > Human
# messages & responses / round	2.31 (0.02)	2.00 (0.00)	2.73 (0.05)	2.32 (0.02)	Human > Hybrid > AI
# asker messages / round	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	-
# AI & human responses / round	1.31 (0.02)	1.00 (0.00)	1.73 (0.05)	1.32 (0.02)	Human > Hybrid > AI
# AI responses / round	0.48 (0.01)	1.00 (0.00)	-	0.37 (0.02)	-
# human responses / round	0.83 (0.03)	-	1.73 (0.05)	0.95 (0.04)	Human > Hybrid
Asker message length (# chars)	14.14 (0.28)	13.08 (0.51)	15.61 (0.56)	14.16 (0.41)	Human > AI
Answerer response length (# chars)	58.91 (1.18)	86.83 (2.60)	46.53 (2.18)	52.96 (1.49)	AI > Human, Hybrid

5.2 Asker Message Type Distribution

Despite that the three types of sessions in our dataset vary significantly in many statistics, we found that the frequency of using message types in the sessions are mostly consistent, with noticeable differences only in a few message types. Table 12 reports the percentage of each message type in the three types of sessions. We group some message types because they appeared a minimal number of times in our dataset—we group all

four types of Repeat and Rephrase message types together as *REP*, and we group Correction, Nudge, and Feedback as *OTHER*.

First, our results show that the frequency of using different types of messages is highly consistent across sessions with various representatives. We compared the overall distribution of message types in the three types of sessions using a Kruskal-Wallis H test. The test results suggest no significant differences between any of the session types regardless of using the original 15 message types ($p = 0.191$) or the grouped 10 types ($p = 0.537$). This indicates that 1) our message classification scheme is highly generalizable and can be applied to different types of sessions, and 2) the use of different message types seems relatively stable when communicating with AI and human representatives (though it is unclear whether the results would remain the same if the customers were told which type of representatives they were talking with).

Table 12: Distribution of user message categories in sessions with AI, human, and hybrid representatives. For each message type, we test significant differences of three session types using the Chi-square test with Bonferroni correction.

Asker Message Type	ALL	AI-only	Human-only	Hybrid	*P* < 0.05 Differences
Q (new query)	35.5%	38.9%	48.1%	28.0%	Hybrid < AI < Human
FQS (follow-up query, self)	5.3%	5.4%	4.8%	5.5%	-
FQO (follow-up query, the other person)	14.5%	13.3%	14.2%	15.4%	-
REP (repeat & rephrase)	5.1%	6.8%	2.9%	5.0%	-
Repeat (repeat without any change)	0.9%	0.7%	1.0%	1.0%	-
Rephrase (rephrase; wording difference)	1.8%	3.2%	0.0%	1.8%	Human < AI
RephraseAdd (rephrase; added some content)	1.8%	2.0%	1.3%	1.9%	-
RephraseDel (rephrase; removed some content)	0.5%	0.9%	0.6%	0.3%	-
CLR (clarify)	4.1%	3.4%	4.2%	4.4%	-
ANS (answer)	3.2%	0.2%	4.5%	4.4%	AI < Human, Hybrid
CH (chit-chat)	16.2%	15.8%	13.2%	17.6%	-
BG (background information)	6.6%	7.9%	1.6%	7.9%	Human < AI, Hybrid
SUP (supplementary information)	7.6%	6.3%	4.8%	9.5%	Human < Hybrid
OTHER (other types)	2.0%	1.8%	1.6%	2.3%	-
Correction	0.9%	0.7%	0.3%	1.2%	-
Nudge	0.3%	0.2%	0.6%	0.1%	-
Feedback	0.9%	0.9%	0.6%	1.0%	-

Kruskal-Wallis H test (H_0: message category distributions in AI-only, Human-only, and Hybrid sessions are not significantly different): $P = 0.537$

Second, results in Table 12 also disclosed the popularity of the message types in conversations. In all three types of sessions, Q, FQS, and CH remain the top three most popular types. About half of the messages in the sessions are directly asking new questions (Q, FQS, and FQO make up 55% of all messages), with over 1/3 are follow-up questions (19.8% out of 55%). Also, 18.3% of the messages' purpose is to provide contexts and details (BG, SUP, ANS, and Correction). Askers also used 16.5% of the messages (Chit-chat and Nudge) to maintain conversations. In contrast, restating questions (REP) and understanding answerers' responses (CLR and Feedback) take up only 5.1% and 5% of the total. These statistics provide important guidelines for designing chatbots that can better respond to different types of user messages.

Third, the message type distribution also suggests that a substantial number of messages in a conversation session are closely linked with some other messages/responses in the same session. By our definition, FQS, FQO, REP, CLR, ANS, BG, SUP, and Correction messages all have related messages or responses. They take

up 47.1% of all askers' messages. This shows that the messages and responses in a conversation are highly related to each other, suggesting that the importance of modeling context information in designing chatbots.

Forth, we did have also observed significant differences for a few specific message types across different session types. Some of the differences are related to the settings of the customer service system, e.g., we had only observed one case of Answer in AI-only sessions because the AI chatbot does not provide questions as responses. For other differences, we suspect one possible reason lies in the question routing strategy in the customer service system—for example, if the system has higher confidence scores for a certain message type, those messages will be less likely routed to human representatives and thus will have a lower percentage in the human-only sessions. However, it returns further study to understand the differences of using messages in sessions with different representatives.

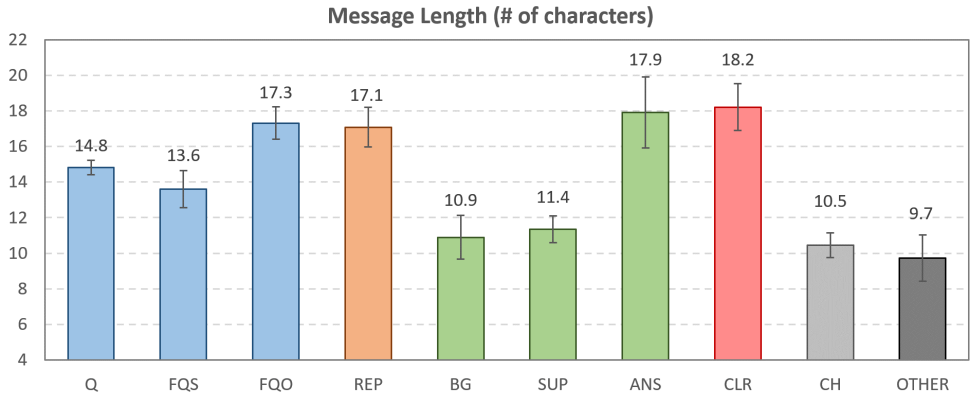

Figure 3: Length of messages classified into different types (by the number of Chinese characters).

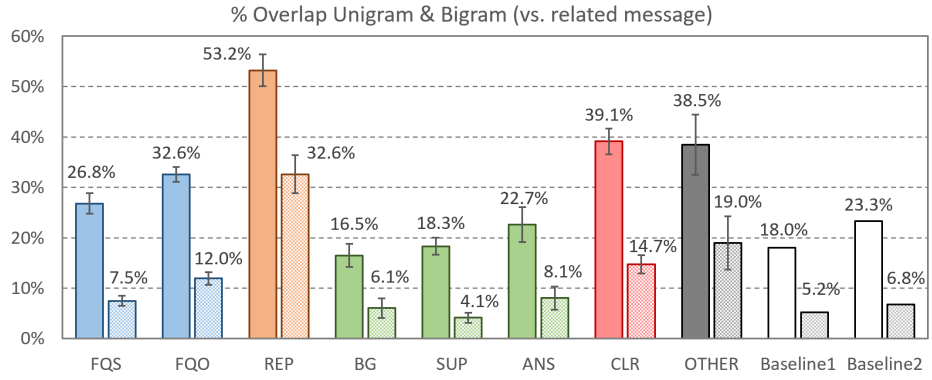

Figure 4: The percentage of overlap unigrams and bigrams (pattern-filled bars) between a message and its identified related message/response. Baseline 1 is the average overlap with a most recent asker message. Baseline 2 is the average overlap with a most recent answer response.

5.3 Characteristics of Asker Message Types

After examining the content of asker messages, we found that different types of messages vary greatly in length (Figure 3) and their similarities with the identified related messages/responses in the session (Figure 4). Here,

we measure message length by the total number of Chinese characters. We measure the similarity of an asker message with its related message/response by the percentage of the common content (by Chinese character unigrams or bigrams) among the asker message itself.

Note that a Chinese word typically includes one, two, or three characters, where we can roughly equivalate a Chinese character to a word root or stem in English. The whole Chinese character set includes over 50,000 different characters, with about 3,500 frequently used ones. We did not examine messages by words because we found that out-of-the-box word segmentation tools[1] did not work well on our dataset (probably because the text messages are noisy and used many verbal expressions).

Figure 3 shows that some types of messages are much longer than others. Particularly, FQO, REP, ANS, and CLR messages are significantly longer than BG, SUP, CH, and OTHER messages in our dataset by to a Tukey HSD post hoc test (the difference of each pair is at least significant at 0.05 level).

Figure 4 reports the percentage of overlap character unigrams (solid color bars) and bigrams (pattern-filled bars) between an asker message and its identified related message/response in the dataset. We did not report results for Q and CH because they do not have related messages/responses. Figure 4 also shows the overlap values with a most recent asker message (baseline 1) and a most recent answerer response (baseline 2) across the whole dataset for comparison. The results for overlap unigrams and bigrams are mostly consistent.

Results show that, except BG, SUP, and ANS, other types of asker messages share much more common contents with their identified messages or responses than two random adjacent messages/responses (baseline 1 and 2). This also demonstrates that the manually labeled related messages/responses are probably accurate.

Results also show that overlap percentages vary a lot in different message types. The overlap unigrams take up over 50% of the content in REP messages, 30%–40% of the content in FQO, CLR, and OTHER messages, and lower proportions in other message types. The highest percentage of overlap content in REP messages is not surprising since the intention of the REP messages is to restate their related messages. The high overlap content percentages in FQO and CLR messages are probably because the askers need to refer to the overlap content when asking follow-up questions or clarifying previous responses.

To conclude, results in this section show that different message types vary in characteristics related to their contents, which provides potential opportunities to recognize message types automatically based on contents.

5.4 Message Type Transition: After a Q Message

We further examine the use of different message types in a contextual manner—such as right after or before a message type. Figure 5 illustrates the transition probabilities to different message types after a Q message, i.e., the chances of having different message types if the previous asker message is Q. We also calculate the chance that the Q message is the last asker message in the session (Q→END). We separately examine each individual session types and all sessions. We focus on Q messages because it is the most common message type.

Figure 5 shows that the use of different message patterns right after a Q message is mostly consistent across different session types with some differences. The top three most frequent message types after Q is Q, FQO, and END (which means that the Q is the last asker message of the session). Also, CH, SUP, FQS, and REP are also relatively frequent types. This suggests that the main pattern of an informational session is to keep on

[1] The Chinese writing system does not put a white space between words such that we need to use NLP tools to segment words from text.

asking questions (including follow-up questions), occasionally with other messages to refine and complement the questions or clarifying received responses.

We have also observed some differences of message type transition in different sessions. Particularly, we noticed that the chances of Q→REP is much lower in human-only sessions, indicating that human-only sessions probably have better response quality (such that askers do not need to restate the same needs multiple times). To conclude, these results indicate the possibility that the usage of message types may be related to the type of representatives in a session, but it requires further study to verify due to our quasi-experimental design.

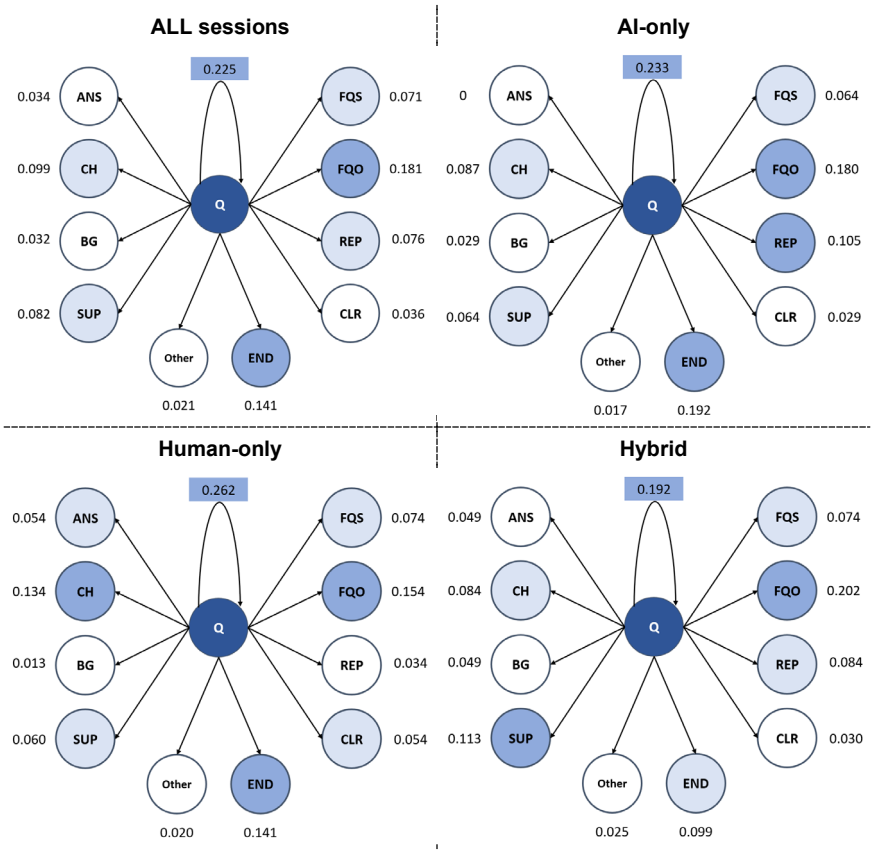

Figure 5: User message category transition probabilities after a Q message (Q→?) in different types of sessions.

6 CONCLUSION

In this paper, we introduced a fine-grained hierarchical classification scheme for asker messages in online text-based informational conversations. Our scheme included five categories of messages (15 specific types) linked to three higher-level asker intentions: describing information needs, understanding information, and maintaining conversation. The detailed message types share some similarities with previous studies of query reformulation patterns and dialogue acts, but we put a special focus on the function of the message for assisting users during the conversation to acquire information. We have also examined the annotation results on a real-world dataset and reported statistics comparing different message types and in sessions with different representatives.

We make the following contribution in this study:

First, we have concluded and introduced a classification scheme for categorizing asker messages in online informational conversation. To the best of our knowledge, this is the *first* classification scheme for online text-based informational conversations, which shed lights on understanding a wide range of real-world applications, especially online text-based customer services. We acknowledge that many previous studies examined user request patterns in search engines, conversational dialog systems, and intelligent personal assistants, but our study and scheme is novel from two aspects: 1) informational conversation shares some similarities with but is very different from these applications, e.g., it provides more interactive communication and direct access to the information compared with a search engine, it focuses on information seeking and acquisition tasks compared with dialog systems and intelligent assistants; 2) we design our classification scheme from a novel aspect and link the message categories with higher-level user intentions for information acquisition.

We have also demonstrated that the classification scheme is practical and actionable. As we described, we have successfully annotated a real-world dataset with highly specialized conversation topics (medical health). Our annotators do not have any prior knowledge related to this specialized topic, but they had still come to very reasonable agreements during the annotation. We acknowledge that the messages in our data is in Chinese, but our classification scheme does not include rules or details related to the language used in the conversation. This suggests that the proposed scheme is likely to generalize to other scenarios and by lay people.

We believe our novel classification scheme provides significant guidance on future work related to online informational conversation, including both work for understanding human factors and designing systems and interactive techniques. For example, researchers may apply our scheme to annotate and examine informational conversations, and systems may design techniques for classifying user message types, recognizing related messages, and prepare specialized responses accordingly in the future.

Second, we have also presented detailed comparisons among different message types. The results provide insights to understand different types of asker messages in a session. Particularly, we have observed that many types of messages vary significantly in content characteristics, such as length and their similarities with other messages in session. On the one hand, this provides a second look into the validity of our classification scheme and data annotation consistency, because many of the observed differences can be explained well based on the definition of our message types (e.g., most message types have high content similarities with their related messages or responses). On the other hand, this offers clues to design techniques to recognize message types automatically—for example, message length and content similarity with previous messages & responses may be effective features for automatically classifying asker message types.

Third, we have also presented an initial exploration of the possible relations between asker message types and sessions with different types of representatives (AI-only, human-only, and hybrid). Our initial observation is that the use of message types are quite stable across sessions of different types of representatives, but we did also observed that the usage frequency for some specific message types (such as Q, REP, BG, and SUP) can be significantly different in different types of sessions. This provides a basis for further studies to examine the relationship between different types of representatives and users—we believe this is a fundamental research question in online informational conversations as it is more and more common to offer hybrid chat services.

As the first study of this topic, our work also has some limitations. We leave them for future work. First, we acknowledge that our dataset included only a very specialized topic (health insurance) and is in Chinese. We advise future work to further verify the generalizability of our scheme and findings (though we believe that our

study can easily be replicated in other languages, e.g., one can examine word-based unigram and bigram overlap in English language datasets). Second, our selection of sessions with different types of representatives are quasi-experiment, and we are aware that the selection may be affected by some message characteristics (i.e., the chatbot's confidence for responding to these messages). Thus, we also suggest further studies to have randomly assigned experiments to examine the relationship between the types of representatives and the use of messages.

REFERENCES

[1] Core, M. G., & Allen, J. (1997, November). Coding dialogs with the DAMSL annotation scheme. In AAAI fall symposium on communicative action in humans and machines (Vol. 56, pp. 28-35).

[2] Jurafsky, D., Shriberg, E., & Biasca, D. (1997). Switchboard-DAMSL labeling project coder's manual. Technická Zpráva, 97-02.

[3] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics, 26(3), 339-373.

[4] Walker, M., & Passonneau, R. J. (2001). DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. In Proceedings of the First International Conference on Human Language Technology Research.

[5] Ferschke, O., Gurevych, I., & Chebotar, Y. (2012, April). Behind the article: Recognizing dialog acts in Wikipedia talk pages. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 777-786). Association for Computational Linguistics.

[6] Oraby, S., Gundecha, P., Mahmud, J., Bhuiyan, M., & Akkiraju, R. (2017, March). " How May I Help You?" Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts. In Proceedings of the 22nd international conference on intelligent user interfaces (pp. 343-355).

[7] Braslavski, P., Savenkov, D., Agichtein, E., & Dubatovka, A. (2017, March). What do you mean exactly? Analyzing clarification questions in CQA. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (pp. 345-348).

[8] Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015, May). Automatic online evaluation of intelligent assistants. In Proceedings of the 24th International Conference on World Wide Web (pp. 506-516).

[9] Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. Computer Speech & Language, 24(2), 150-174.

[10] Traum, D. R. (2000). 20 questions on dialogue act taxonomies. Journal of semantics, 17(1), 7-30.

[11] Hassan Awadallah, A., Gurunath Kulkarni, R., Ozertem, U., & Jones, R. (2015, October). Characterizing and predicting voice query reformulation. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 543-552).

[12] Jiang, J., Jeng, W., & He, D. (2013, July). How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 143-152).

[13] Anick, P. (2003, July). Using terminological feedback for web search refinement: a log-based study. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 88-95).

[14] Bruza, P., & Dennis, S. (1997, June). Query Reformulation on the Internet: Empirical Data and the Hyperindex Search Engine. In RIAO (Vol. 97, pp. 488-499).

[15] Dang, V., & Croft, B. W. (2010, February). Query reformulation using anchor text. In Proceedings of the third ACM international conference on Web search and data mining (pp. 41-50).

[16] Guan, D., Zhang, S., & Yang, H. (2013, July). Utilizing query change for session search. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 453-462).

[17] Huang, J., & Efthimiadis, E. N. (2009, November). Analyzing and evaluating query reformulation strategies in web search logs. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 77-86).

[18] Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during Web searching. Journal of the American society for information science and technology, 60(7), 1358-1371.

[19] Kelly, D., Gyllstrom, K., & Bailey, E. W. (2009, July). A comparison of query and term suggestion features for interactive searching. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 371-378).

[20] Rieh, S. Y. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management, 42(3), 751-768.

[21] Yue, Z., Han, S., He, D., & Jiang, J. (2014). Influences on query reformulation in collaborative web search. Computer, 47(3), 46-53.

[22] Shokouhi, M., Jones, R., Ozertem, U., Raghunathan, K., & Diaz, F. (2014, July). Mobile query reformulations. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 1011-1014).

[23] Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. arXiv

preprint cmp-lg/9704004.

[24] Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: selection and effectiveness of search terms. Journal of the American Society for Information Science, 48(8), 741-761.

[25] Koenemann, J., & Belkin, N. J. (1996, April). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 205-212).

APPENDIX: ORIGINAL MESSAGES IN CHINESE (MANDARIN)

Table A1: An example Chinese message classified as **Follow-up Question, Self (FQS)**.

Q	也就是说一年没有一万 X 保不赔是吗？
	对的。
FQS	*一万以内就去社保或者其他医疗保险报销是吗？*
	是的。

Table A2: An example Chinese message classified as **Follow-up Question, Answerer (FQA)**.

Q	请问下住院后，医药费是自己先出吗？
	您好，咱们这个是医疗报销型保障，需要您自行垫付，后续拿理赔资料向保险公司提出理赔申请。
FQA	*后续要多久？*
	您好，没能明白您的意思。

Table A3: An example Chinese message classified as **Rephrase**.

Q	交的保费以后还返还吗
	X 保产品是一款高杠杆的消费型产品，最低仅须百余元即可享有高达 600 万的保障，您可以提供下点击点击小程序——健康险——测算首年保费，可以查看不同年龄的保费。
Rephrase	*我问保费以后还能退不*

Table A4: An example Chinese message classified as **Rephrase, Add**.

Q	住院了能保吗
	X 保是一款医疗报销型产品，主要报销医疗费用包括：住院医疗费用、特殊门诊医疗费用、门诊手术医疗费用、住院前后门急诊医疗费用。因此，不仅限于住院医疗费用。
Rephrase, Add	*还没有保。现在住院了能保吗*

Table A5: An example Chinese message classified as **Rephrase, Delete**.

Q	3 岁多孩子，应该选无社保吧？
	请您稍等，正在为您描述。
	您好！本产品有无社保身份均可参保，但参保保费会有所不同
Rephrase, Delete	*孩子都是无社保的吧？*

Table A6: An example Chinese message classified as **Background (BG)**.

Background	*我去年骨折住过院*
	如因下列原因在两年内住院，则为例外事项，可进行投保：A)分娩；B)急性呼吸系统疾病；C)急性胃肠炎、急性阑尾炎；D)胆结石经治疗后 2 年内未复发；E)胆囊息肉已手术切除且病理结果为良性；F)意外住院在 5 天以内且已痊愈，并无后遗症或器官缺损。
Q	是否不满足条件

Table A7: Example Chinese messages classified as **Supplement (SUP)** and **Answer (ANS)**.

Q	请问这个保费是每年递增吗？

	随着年龄的增长，发生意外或疾病的风险会越来越高，保费自然也会越来越高。但 X 保本身是一款普惠型的产品，即便是 60 岁的被保险人，年保费也仅 1000 多元，即每月 100 多元即可享受高达 600 万的医疗费用保障，是人人可负担的高性价比产品。
Supplement	*我今年 **49** 岁*
	请您稍等，正在为您描述。
	请问您有无社保呢？[玫瑰]
Answer	*有*
	您好，目前 49 周岁有社保 858 元哟

Table A8: An example Chinese message classified as **Clarify (CLR)**.

Q	X 保买了，多久期间能用
	本产品投保后次日零时生效，等待期为 30 天。
	等待期又称观察期，是指客户投保后的 30 天期间，在这个期间内客户因为疾病发生保险事故无法获得赔付。等待期的设置是为了保障健康客户的利益，防止客户带病投保，保障大部分投保人的利益。
CLR	*30 天过后才能用*
	是的。

Table A9: An example Chinese message classified as **Feedback**.

Q	哦哦 从微信端投保成功后的次日零时开始生效 这句话的意思我可以理解为今天从微信端投保成功后 第二如果我有什么重大疾病就可以理赔吗？
	您好，客户投保后的 30 天期间，在这个期间内客户因为疾病发生保险事故无法获得赔付的。
	但是意外是没有等待期的。[愉快]
Feedback	*看不懂 能解释通俗点吗？*
	请您稍等，正在为您描述。
	简单地说就是：投保成功后，因为意外导致的保险事故没有等待期，但是因为疾病住院的话有 30 天的等待期，也就是您说的观察期。[愉快]
	这么解释您还清楚吧。