Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Medieninformatik

Building Complex Queries in Conversational Search

Bachelor's Thesis

Xiaoni Cai

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Jan Ehlers

Submission date: September 18, 2020

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, September 18, 2020

. . . .
Xiaoni Cai

Abstract

> I think more and more, information retrieval is moving from helping people find information, to helping people get things done.
>
> — Susan Dumais[1]

In a traditional search interface, seekers can quickly modify their query by adding, removing, or replacing terms in the text input field. In a conversational search interface, on the other hand, there is no input field that keeps the last query for quick modifications. So how can seekers modify the queries instead? A straightforward solution is to repeat the entire query with modification, but this is tedious and increases the possibility of system errors due to the complexity of recognition. To solve this problem, this thesis proposes a Query Rewriting Layer. This layer implements the translation of a sequence of seeker utterances into queries of a traditional query language. By referring to the context of the whole session, the Query Rewriting Layer supports seekers to pose complex queries in a multi-turn conversational search. To inform the design of the layer, we perform a large-scale behavior analysis using crowdsourcing, which provides insight into how seekers reformulate their queries while interacting with systems and into the ambiguities of their requests. Based on this analysis, we implement a prototype front-end of the Query Rewriting Layer as a proof-of-concept. For the prototype back-end, we present the idea of taking advantage of the search engine Elasticsearch to efficiently implement the utterance translation and query rewriting.

[1] MSR Podcast, September 18, 2019. `https://www.microsoft.com/en-us/research/podcast/hci-ir-and-the-search-for-better-search-with-dr-susan-dumais/`

Contents

Chapter 1

Introduction

Using the traditional web search engine is a ubiquitous manner for seekers to search and it is mainly based on desktop and mobile operating systems. Seekers enter a text query outlining the information they need [Kaushik et al., 2020]. The information retrieval (IR) system locates items related to seeker utterance in a way akin to manual library-based approaches of acquiring, indexing, and searching information but is far more efficient [Sanderson and Croft, 2012].

Searching for information on the Web suffers from some limitations. Seekers might have difficulty using the search engine query language to query a search engine [Cabanac et al., 2008]. It is challenging for the traditional web search system to satisfy exploratory and open-ended information needs of seekers, especially when they are not familiar with the domain of question [Eickhoff et al., 2014]. Additionally, the search engine has to return an answer based on indices, combining information from external knowledge bases [Kenter and de Rijke, 2017]. To overcome these restrictions, an alternative search paradigm, conversational search, which a seeker and a natural-language-based system are able to engage in a dialogue, came into view.

In 1960, Licklider posed a question in "Man-Computer Symbiosis" paper [Licklider, 1960],

> `How "desirable" and "feasible" speech communication between human operators and computing machines could be?`

With the development of spoken language technologies such as Automated Speech Recognition (ASR) and Natural Language Understanding (NLU), speech-based conversational search using smart speakers like Google Home and Amazon Echo is increasingly integrating into daily life. From 2017 to 2019, the number of smart home devices compatible with Amazon's virtual assistant

Alexa has increased intensively from 4,000 to 60,000 [Statista, 2020]. Ammari et al. [2019] found that Search or information queries was one of the most prevalent uses of Google Home (at 26%) and of Amazon Alexa (at 19.4%) by analyzing the commands in users' device usage logs.

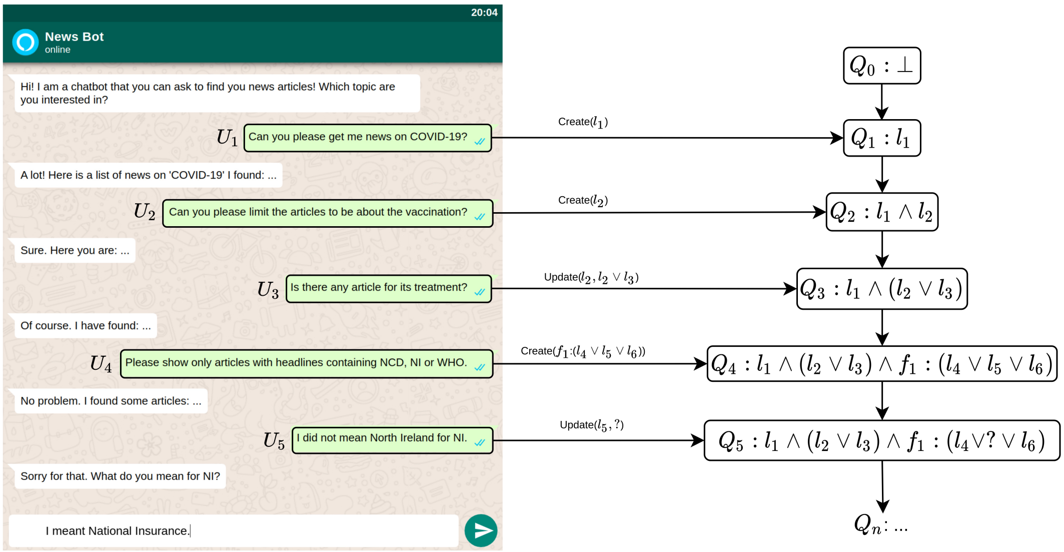
Besides voice-only interaction, text-based chatbots are also widely spreading in the area of conversational search, which are designed to interact with seekers in dialogues using natural language. Chatbots search for information on a variety of topics like news, restaurants and the weather. Seekers engage with these chatbots back and forth by sending certain utterances, answering chatbot's follow-up questions and interacting with the results provided by chatbots [Avula et al., 2018]. Users are easily inured to conversational search whether it is based upon speech or text as conversational search imitates the way humans engage with each other and is intuitively attractive [Kaushik et al., 2020].

Nevertheless, the conversational search system has difficulty in understanding seekers' conversations for the reason that the seeker utterances can be truncated, colloquial and contextually dependent and commonly face coreference and omission problems [Lin et al., 2020].

How can a compelling conversational search system be expected? Fraser et al. [1998] indicated that humans interact with the computer system on a turn-by-turn basis and natural language plays an imperative role in communication. In a natural language conversation search, the information involved in the earlier utterances can be referenced, even if it is implicit [Kenter and de Rijke, 2017]. In other words, the system should keep track of the previous requests made by seekers over multi-turn conversational interactions and meanwhile, capture relevant context that is crucial to resolve the seeker's current information needs.

From seekers' perspective, such a conversational search system enables seekers to continuously develop their query by clarifying cumulatively (e.g., adding more items or deleting) and satisfies seekers with more complex information needs. Seekers are allowed to refer to previous discussions but omit already mentioned concepts and assume implicit context during the conversation [Yu et al., 2020]. More than that, they can refer to past statements explicitly, for instance, to argue which statement is incorrect and misunderstood by the system, or inquiry the forgotten queries [Radlinski and Craswell, 2017].

Hence, conversational search is commonly considered as one of the most important topics in information retrieval [Culpepper et al., 2018]. The core of our thesis is to enable seekers to pose complex queries stepwise during the course of a multi-turn conversational search. We first explore relevant researches in four fields of conversational search, query formulation, query languages, and query rewriting (Chapter 2). Based on these pieces of knowledge, we conceptualize the query reformulation under the principle of Create, Read, Update, and Delete (CRUD) operations (Chapter 3). In the same Chapter, we advance a Query Rewriting Layer, in theory, to provide a structured model of building complex queries by rewriting the existing query recursively, as shown in Figure 1.1.



Hi! I am a chatbot that you can ask to find you news articles! Which topic are you interested in?

U_1 Can you please get me news on COVID-19?

A lot! Here is a list of news on 'COVID-19' I found: ...

U_2 Can you please limit the articles to be about the vaccination?

Sure. Here you are: ...

U_3 Is there any article for its treatment?

Of course. I have found: ...

U_4 Please show only articles with headlines containing NCD, NI or WHO.

No problem. I found some articles: ...

U_5 I did not mean North Ireland for NI.

Sorry for that. What do you mean for NI?

I meant National Insurance.

Create(l_1)
Create(l_2)
Update($l_2, l_2 \vee l_3$)
Create(f_1:($l_4 \vee l_5 \vee l_6$))
Update(l_5, ?)
$Q_0 : \perp$
$Q_1 : l_1$
$Q_2 : l_1 \wedge l_2$
$Q_3 : l_1 \wedge (l_2 \vee l_3)$
$Q_4 : l_1 \wedge (l_2 \vee l_3) \wedge f_1 : (l_4 \vee l_5 \vee l_6)$
$Q_5 : l_1 \wedge (l_2 \vee l_3) \wedge f_1 : (l_4 \vee ? \vee l_6)$
Q_n: ...

Figure 1.1: Interactive interface between participants and chatbot used to collect seeker utterances with partial modifications and the corresponding query $Q_0, ..., Q_n$ after rewriting. Each Q_i consists of Boolean clauses except Q_0, which is empty and denoted by $\perp$.

Subsequently, in Chapter 4, in order to have a better understanding of seeker interaction and utterance intent, we recruit participants via the Mechanical Turk marketplace, who are instructed to interact with a "chatbot" using a conversational text-based search interface (Figure 1.1) to conduct a topic-oriented search study comprised of 12 predefined tasks for different reformulation intents. Our contribution is to collect and review their utterances to recognize the templates of utterance. Combining with the obtained seeker utterance templates, in practice, we design a conversational query interaction model as the prototype front-end intended to prove the proposed Query Rewriting Layer concept in Chapter 5. For the prototype back-end, we propose the idea of utilizing the search engine Elasticsearch to implement the translation of utterances and query rewriting. A cross-evaluation method is used to evaluate the generalizability of templates. In Chapter 6, we analyze the collected reformulation patterns and discuss the detected ambiguities that occurred in the reformulation of the query. Finally, we present our conclusions and outline future work in Chapter 7.

Chapter 2

Background

In this chapter, we introduce the related background for this thesis and organize it into four sections. We start by presenting the research related to conversational search and highlight the concepts closely linked with our study in Section 2.1. The characteristics and classifications of seeker utterances in the conversational search are described in Section 2.2, revealing the way in which seeker formulates and reformulates their query. Next, in Section 2.3, we briefly introduce the features of the query language of the search engine Elasticsearch. We then present tense relevant studies on query rewriting and compare them with our study in Section 2.4.

2.1 Conversational Search

In a conversational search, seekers interact with the information system using natural language with multi-round interactions in order to satisfy an information need. Human-like communication in human-computer interaction can be traced back to the notion of 'Man-machine symbiosis' proposed by the Licklider [Licklider, 1960]. In recent years, due to the advancement in Artificial Intelligence technology and increasing popularity of conversational agents (e.g., chatbots, smart speakers), a wide range of relevant researches are deployed and a variety of working systems are being developed, which brought the fantasy of science fiction to real life. Such a system with either a text-based or voice-based interface is capable of answering a question or seeking information.

Most of the previous work is targeted towards the single-turn conversation. That is to say, the system only takes the current utterance into account to rank results. However, in contrast to that, several researchers contribute with a multi-turn conversational search system, which keeps track of the previous utterances as contextual information to select a response for the current message. To enable the system to naturally and efficiently fulfill various information needs of seekers over multiple rounds of conversation, Radlinski and Craswell [2017] has advanced five desired properties of a conversational search system.

User Revealment: The system elicits seekers' actual needs and assists them in formulating it properly.

System Revealment: The system informs seekers which functions it can support or not provide during a conversation. (e.g., "*tell me what kind of news you want to hear.*")

Mixed Initiative: Both the system and seekers can take the initiative whenever appropriate. The system takes the initiative to clarify user's information need at some points, such as "*Do you mean. . . ?*", in turn, the user takes the initiative to drive the conversation at other times.

Memory: Seekers can refer to earlier statements. (e.g., correct their own utterance using "*what I mean is. . .* ")

Set Retrieval: The system has the ability to infer the utility of sets of complementary items.

They argued that the more complex the search task, the more valuable a back-and-forth conversation. We also reference some of the scenarios while designing our user study. For instance, **Multi-Item Faceted Elicitation**, the seeker searches for a set of items. The system needs not only to estimate the utility of every single item but also to combine the utilities of multiple items to reach an assessment of an entire set. In our case, seekers inquire about a list of trips consisting of a specific destination, different transportation options, hotel requirements, and sightseeing.

In such conversation settings with the aforementioned properties, users do not need to repeat already mentioned concepts in previous turns throughout the conversation. The system asks clarifying questions and elicits information from user utterance in each simple turn, cumulatively describing a complex information need. Through cumulative clarifications, the system tends to move closer to the user's goal.

Radlinski and Craswell [2017] also presented a conversation action space in their conversational search model to summarize interaction patterns between system and user, representing possible feedback from the system and expected responses from the user. These interaction patterns inspired us which designing the tasks of the study with different formulation intents. For example,

Partial Item System - Pref/Rating User: A user provides partial information that can be matched items in various ways. The conversational system confirms a slot that has been inferred, such as "*Do you mean National Insurance?*"

Partial Item System - Critique User: A user indicates specific individual facet values. For the prompt "*Do you mean National Insurance?*", instead of answering a simple yes/no, the user may reply "*no, I mean North Ireland.*"

2.2 Conversational Query Formulation

Lin et al. [2020] summarized observations about the characteristics of conversational seeker utterances. First, A session orients around a main-topic and the subsequent turns delve further into multiple subtopics throughout the session, however, each of which only lasts a few turns. Second, they classified the ambiguity degree of seeker utterances into three categories. The first category comprises utterances with clear implications. The second category includes those starting a subtopic, and the last category consists of most ambiguous utterances that continue a subtopic.

Walboomers and Hauff [2020] divided seeker utterances into two groups: natural language utterance (**NatLang**) or keyword (**Keyword**). **NatLang** incorporates all utterances that seekers will naturally verbalize it in a conversation. For instance, *Hi! Could you please show me all articles about COVID-19?* **Keyword** represents those utterances where seekers conceptualize their information need concisely and formulate it as a keyword query to the chabot. An example could be: *COVID-19 articles.*

In addition, they also distinguished between non-querying messages (**NonQ**) and querying messages (**Query**). **NonQ** involves all utterances without the intention of seeking information. On the contrary, they are informing utterances, consisting of greeting to the chatbot, thanking the chatbot, giving positive or negative feedback to the chatbot, or any other utterances that do not directly convey the information need. **Query** includes all utterances for the purpose of satisfying a certain information need.

To develop a conversational search system, it is crucial to understand the way how seekers formulate and reformulate their utterances during the dialogue interaction. There are a large number of researches on query reformulation in traditional information retrieval systems. Their query reformulation patterns can be partially applied to the conversational search system. However, they missed how the conversational system can fuse context information into seeker utterances to fulfill a cumulative information need.

In the context of text-based conversational search, Qu et al. [2018] proposed a query reformulation taxonomy by finding patterns that appear frequently in user intent. Walboomers and Hauff [2020] has classified query reformulation into four types: (near-exact) duplicate (**Dup**), rephrase information need (**Rehp**), a new information need about a familiar topic (**NewInfNd**) and topic switch (**TopS**).

Apart from the text-based search interface, typing errors do not exist in a voice-based search. However, there are system recognition errors (e.g., missing words, incorrect words) and system interruptions by interacting with a voice-only search interface [Jiang et al., 2013].

Sa and Yuan [2020] conducted a Wizard of Oz user experiment, and the results revealed the following: (1) Seekers prefer implementing partial query modification rather than speaking the entire query in voice search. 40.8% (191 of 468) spoken queries were complete modification in contrast to 59.2% (277 of 468) partial modification. (2) The query modification type **Specification** (as 'Adding' in traditional textual systems) was most widely used, and **Generalization** (Deleting) was more commonly used than other remaining types. (3) The most frequently used strategies in partial query modification were: **specific operation**, **partial repeat**, and **appending**. When conducting **specific operation**, the most commonly used operation commands were the ones that replace terms.

2.3 The Elasticsearch Query Language

Information Retrieval Query Languages are computer languages used to build queries with the intent to locate documents, including information related to certain areas of inquiry.

Elasticsearch is a distributed, horizontally-scalable, real-time and multi-tenant textual search engine built on Apache Lucene as back-end [Mu et al., 2019]. Elasticsearch has a high performance in Data Searching and is able to detect the data structure and data types automatically.

A human-entered text or a text translated from spoken language is interpreted into a Query. The language in a query string provided in Elasticsearch has the following features[1]:

1. A Query is broken up into operators and terms.
2. There are two types of terms: Single Terms and Phrases. A Single Term is a single word. A Phrase is a collection of terms enclosed by double quotes.
3. Multiple terms can be combined with Boolean operators such as 'AND' or 'OR' to form a more complex query.
4. It supports searching any field by specifying the field name followed by a colon ":" and then the specific term you are looking for.
5. It supports single and multiple character wildcard (e.g., '*', '?') searches within single terms.
6. Tokenizers are functions for transforming strings. Analyzer is a set of one or more tokenizers or filters. Filters apply to every token returned by the tokenizer.

More than that, Elasticsearch can translate multiple queries into a recursive listing and save all kinds of I/O operations while looking for data, which brings us the benefit of building a complex Query to satisfy a cumulative information need.

2.4 Query Rewriting

Conversational Query Rewriting aims to reformulate ambiguous utterances that depend on previous turns in dialogue into unambiguous utterances independent of conversational context. We explore existing studies on this topic and compare them with our study.

Kellar et al. [2007] divided information-seeking tasks into two types: Fact finding and Information gathering. **Fact finding**: seekers attempt to look for

[1] https://lucene.apache.org/core/2_9_4/queryparsersyntax.html

specific facts or pieces of information. **Information gathering**: seekers ask for the collection of information and there is no specific answer. Fact Finding also always refers to question answering. The seeker asks a question and the system finds a specific answer as the response.

Several studies present approaches and evaluations on conversational query rewriting under the scenario: conversational question answering. Vakulenko et al. [2020] stated that there are two variants of the conversational question answering distinguished by the expected answer types. **Retrieval Conversational Question Answering**: the system gives a ranked list of documents from the collection as a response to a natural-language question. **Extractive Conversational Question Answering**: the system returns a text span extracted from a document to answer a natural-language question. They proposed two sub-models for the conversational question answering system. Model Question Rewriting (QR) is in charge of handling contextual information of an input question. As a result, the QR model produces an explicit question, which is interpretable outside of the conversation and equivalent to the input question. Then the system inputs the generated explicit question to a standard Question Answering Model (QA), which is pretrained by non-conversational datasets. QA processes the explicit question and returns the corresponding answer. They evaluated the performance of the system to tackle both retrieval and extractive question answering tasks.

Yu et al. [2020] developed two methods based on rules and self-supervised learning to generate weak supervision data. In this way, they convert large numbers of ad hoc search sessions into ambiguous, context-based queries. Then they utilized these data to fine-tune the GPT-2 rewriter. GPT-2 learns the context dependencies in the conversational search queries. In the end, they evaluated the effectiveness of the fine-tuned GPT-2 for conversational query rewriting.

Similarly, Lin et al. [2020] proposed two query reformulation approaches: historical query expansion (HQE) and neural transfer reformulation (NTR). HQE applies query expansion as a common technique for addressing query reformulation in the traditional information retrieval system. NTR is a transfer learning method that leverages human knowledge of conversational query understanding to train a neural model capable of imitating human behavior to rewrite questions in a dialogue interaction. The difference is that they specialized in conversational passage retrieval (ConvPR) instead of generic conversational query answering.

However, our study is mainly concerned with conversational information seeking scenarios rather than answering questions. We return a list of relevant documents as a response to fulfill the information needs of seekers. In addition to disambiguating queries by referring to the utterances produced earlier in a conversation, we attempt to rewrite multiple queries to build a cumulative information need. On the other hand, the conversational question answering systems in the aforementioned studies are open-domain. We concentrate on how humans interact with chatbot and rewrite utterances in topic-oriented dialogues with specific intents instead.

Chapter 3

Conceptualizing the Query in Conversational Search

In a natural conversational search, in contrast to traditional information retrieval, the utterances produced in previous turns can be an additional source of information that be considered when the system tackles with the seeker utterances. In general, the seeker formulates an utterance to give the system instructions about modifying the existing query. The system interprets a natural language utterance to identify the instructions and take action on the existing query. It is beneficial for the system to interpret utterances in the context of the whole session as the system can build a cumulative understanding of the seeker's information need over a multi-turn conversation.

Unlike the taxonomy of query reformulation for traditional information retrieval, only a few studies present the taxonomy of query reformulation types and patterns under a conversational context with a text-based or speech-based interface. However, we did not aim to propose a comprehensive and systematic taxonomy of query reformulation in conversational search scenarios. Instead, by summarizing previous studies, according to different intents, we simplified the most common operations that can be used to modify the existing query to only four types: Create, Read, Update and Delete, also known as acronym CRUD. The details will be elaborated on in Section 3.1. We propose a Query Rewriting Layer in Section 3.2, demonstrating how the system rewrites existing queries to build more complex queries recursively.

3.1 CRUD Operations on the Query

In computer programming, the acronym CRUD refers to four basic functions, Create, Read, Update and Delete, that can be performed on resources. In our case, the system translates a natural language seeker utterance into a query in a single turn. After translation, the system recognizes the operations, terms, or phrases, which are the components of a query. With this information, the system executes the operations on the existing query generated in previous turns with related terms or phrases. We divide the operations that can be applied to queries into four types: **C**reate, **R**ead, **U**pdate and **D**elete. Besides, we also classified the object of operation (Target in the table below) into three categories: **Query**, **Part**, **Literal**. A **Literal** can be an individual word filter such as "treatment", similar to the concept of a single term in a query. A **Part** is a group of literals such as "vaccination and treatment", similar to the concept of a phrase in a query. A **Query** refers to an entire query.

In the following, we introduce the mechanisms for completing the CRUD operations in conversational search and give some examples.

Create: create or add new entries. (1) In a new session, the seeker makes the first querying message. The system converts the message and creates the first query in the session. (2) The seeker asks for adding a series of restrictions to filter the previously obtained results. The system adds those required filters to the existing query. (3) The seeker requests to add only one condition to filter the results, and the system adds this filter to the existing query. For (2) and (3), the specified conditions can be added to filter the results that have certain words, or these conditions include the negations of certain words, that is to say, filtering out the results that have these words.

Read: read, retrieve, search or view existing entries. (1) The seeker wants to review the existing query after several rounds of modification because of forgetting or any other reason. The system reads out or shows the required query to the seeker. (2) The seeker asks for details about a particular part of the existing query. (3) The seeker inquiries a single term of the existing query.

Update: update or edit existing entries. (1) The seeker would like to start a new search on a different topic, so the system updates the entire existing query to a completely new query. (2) The seeker requires modifying a specific part of the existing query, such as appending a new literal to a specified part, replacing a certain part with another part or with a new literal. (3) The seeker asks the system to replace the certain term of the existing query with another new literal or even a literal with unknown value. For example, the seeker points out the word misunderstood by the system like *I don't mean North Ireland by NI.* Nevertheless, the seeker does not clear up which correct word "*NI*" refers to. So the system should ask the seeker for confirmation. The

Table 3.1: Example utterances in the news scenario for each of the intents, aligned with the standard CRUD terminology.

Operation	**Target**		
	Query	**Part**	**Literal**
Create	*Show me news about COVID-19.*	*Show me news that contain NCD, NI or WHO in the headline.*	*Is there any news for its treatment?*
Read	*What do I have so far?*	*What filters did I set for the headline?*	*What was the last filter?*
Update	*Please start a new search on flu.*	*Remove my criteria for headline and search any news about economy.*	*I don't mean North Ireland by NI.*
Delete	*No, let's start again.*	*Remove the word filters vaccination and treatment.*	*Remove news about treatment.*

example response from the system could be: *What do you mean by NI?*.

Delete: delete, deactivate, or remove existing entries. (1) The seeker is no longer satisfied with the current session and would like to delete all the existing queries. However, the seeker does not clarify the new topic, so the system needs to either go back to the initial result of the first query or confirm the new topic in the next step. (2) The seeker makes a request to remove a particular part of the existing query, such as *Remove word filters vaccination and treatment*, so the system should not filter the results with these two words but bring those subjects that are not related to these two words back to the returned results instead. These subjects do not exist in the previous results. Interestingly, if the seeker says *Remove vaccination and treatment from the results*, the system should discard the subjects containing these two words from the results. "Deactivating filters" and "removing words" are fundamentally different and will cause different results to return. (3) The seeker asks to remove the individual literal, either a filter or a word.

There are more specific example utterances for each of the collocations of the operations and targets under the news scenario in Table 3.1.

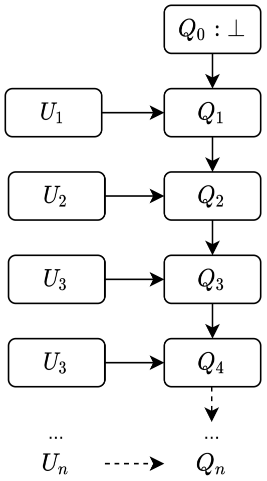


Figure 3.1: a Query Rewriting Layer, where Q_0 is empty, denoted by $\perp$.

3.2 The Conversational Query Rewriting Layer

After the system interprets the seeker utterances and extracts the operations, terms or phrases from utterances, how can the system carry out the operations to rewrite the existing queries generated in earlier turns? We propose a Query Rewriting Layer, which enables the system to modify the existing queries by referring to the context of the whole conversation. As a result, the system constructs more complex queries as the resulting queries.

To be more formal, given a sequence of conversational seeker utterances $U = (U_1, ...U_i..., U_N)$, in a topic-oriented search session S, $S = \{U_1, ...U_i..., U_N\}$. U_i stands for the i-th seeker utterance ($i \in \mathbb{N}^+$) in the session, which is formulated in the turn i. The system is responsible for finding a set of documents for each turn's seeker utterance U_i to satisfy the information need in turn i with the context in previous turns $U_{<i} = (U_1, ..., U_{i-1})$. In turn i, the system will generate a query Q_i, as the result of rewriting existing query Q_{i-1} with seeker utterance U_i by using the query rewriting function ρ. For instance, $Q_0 = \perp$, $Q_1 = \rho(Q_0, U_1)$, $Q_2 = \rho(Q_1, U_2) = \rho(\rho(Q_0, U_1), U_2))$, $Q_3 = \rho(Q_2, U_3) = \rho(\rho(Q_1, U_2), U_3) = \rho(\rho(\rho(Q_0, U_1), U_2), U_3)$, $Q_4 = (Q_3, U_4)$ ... recursively. A structure of Query Rewriting Layer is illustrated in Figure 3.1.

Since seekers formulate their utterances in natural language, it is inevitable to exist some ambiguous sentences in a dialogue, which increase the difficulty of understanding the intents for the system, even the system can reference the contextual information in previous interactions now. For example, in a news-

#	Operation	Target	Arguments	Current query
1	-	-	-	-
2	Create	Query	l_1	l_1
3	Create	Literal	l_2	$l_1 \wedge l_2$
4	Update	Part	$l_2 \rightarrow l_2 \vee l_3$	$l_1 \wedge (l_2 \vee l_3)$
5	Delete	Part	$l_2 \vee l_3$	l_1
6	Create	Part	$f_1 : (l_4 \vee l_5 \vee l_6)$	$l_1 \wedge f_1 : (l_4 \vee l_5 \vee l_6)$
7	Update	Literal	$l_5 \rightarrow ?$	$l_1 \wedge f_1 : (l_4 \vee ? \vee l_6)$
8	Update	Literal	$? \rightarrow l_7$	$l_1 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$
9	Update	Literal	$l_1 \rightarrow l_8$	$l_8 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$
10	Read	Query		$l_8 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$
11	Update	Part	$f_1 : (l_4 \vee l_7 \vee l_6) \rightarrow l_9$	$l_8 \wedge l_9$
12	Update	Query	$q \rightarrow l_{10}$	l_{10}
13	Create	Literal	$\neg l_1$	$l_{10} \wedge \neg l_1$

Table 3.2: The abstracted sequence of operations that each participant has to perform in one of four topic-oriented search sessions: argument, book, news, or trip. "Option 1" is just a test to see whether participants understand the study setting.

oriented search, the system and the seeker engage in a dialogue to satisfy the seeker's information needs with a list of news back and forth. Given the first inquiry, the seeker makes a request: *Can you please get me news on COVID-19?*. It indicates that the system should search for news about COVID-19 (Create Query). However, if the seeker says the next utterance as *Can you please limit the news to be about the vaccination?*, then the question is tricky. Is the seeker more likely to have news on COVID-19 and its vaccination (Create Literal) or ask for news only about vaccination and whether related to COVID-19 or not (Update Query)? Suppose the seeker wants a list of news about COVID-19 and its vaccination, then the seeker says *Are there any news about the treatment?*. The problem is more complicated. There are three interpretations for this utterance with reference to the context in the dialogue: (1) The seeker is likely to have news only about treatment in addition to news on COVID-19 and its vaccination (Create Literal). (2) The seeker is likely to have news only about treatment without any previously acquired results (Update Query). (3) The seeker is likely to have news on COVID-19 and its treatment besides ones on COVID-19 and its vaccination (Update Part).

Indeed, natural language utterances are quite tricky to understand, even harder for the machine. For the above problems, the system can not predict and choose one intent randomly without sufficient conversational query rewriting data to support. It is necessary to undertake a study to understand better how seekers reformulate their utterances with various intents and different operations over a multiple rounds conversation. Additionally, it is essential to observe how seekers distinguish between different intents and whether they confuse or misunderstand some intents. For instance, Create Query and Update Query, Create Literal and Update Part. Furthermore, even for the same operation Remove Part, how do seekers distinguish between "Removing words" and "Deactivating filters" in the expression of utterances?

As stated in the classification of operations that can be taken to modify the existing query in Table 3.1, we design a sequence of operations that seekers will execute in the study in Table 3.2. The study is detailed in the following Chapter 4. Given a set of literals $(l_1, ...l_k..., l_n)$, a query Q_i consists of certain literals as filter conditions, connecting with brackets, $\wedge$ and $\vee$ to join or nest literals. f_1 stands for the function that specifies field name (e.g., 'headline' in the news scenario) with particular values in a query. We include almost all collocations of the operations and targets in the sequence except Delete Query, Delete Literal, Read Part and Read Literal.

Chapter 4

Collecting Human Conversational Query Reformulations

To build a natural and functional conversational search system that can satisfy more complex and cumulative information needs by rewriting the query, we need to understand how seekers engage in such an information-seeking dialogue. Thus, it is necessary to analyze and characterize seeker interaction and utterance intent. In addition, it is also interesting for us to observe how they behave diversely in different topic-oriented search scenarios. So we use crowdsourcing methods to collect conversational query reformulation utterances of the participants under four search scenarios (argument, book, news and trip) in Mechanical Turk.

We introduce how to develop a study with interface and task descriptions by conducting considerable pilot studies in Section 4.1. To gather more qualified utterances in the database, we review the works done by each participant and reject those who attempt to cheat and have over a certain number of "very bad" answers. In the answers of accepted workers, their utterances are labeled as "good" or "bad". The detailed judgment criteria for curation are in Section 4.2. We summarize reformulation patterns in "good" utterances. However, the results of the analysis of collected reformulation patterns will be in part of the independent Chapter 6.

4.1 Crowdsourcing of Conversational Query Reformulation

This section is organized into three subsections. Subsection 4.1.1 includes all specific settings of our study in Mechanical Turk and how to post studies with different scenarios for different countries. To find a reliable version for the formal study, we implement a total of 8 pilot studies in the news scenario with a few participants and analyze the results. The findings and corresponding modifications for the interface and task descriptions are presented in Subsection 4.1.2. The details of all task descriptions and distinct keywords used in task descriptions of different scenarios are shown in Subsection 4.1.3.

4.1.1 Mechanical Turk

We employ workers through Mechanical Turk, a crowdsourcing marketplace that enables requesters to publish their research as HITs. A HIT denotes a human intelligence task and is a question that needs an answer. Worker customers can complete these tasks virtually to get rewards.

There are three qualification requirements to filter reliable and representative workers. The workers are required to (1) have a HIT Approval Rate for all Requesters' HITs of 95% or higher, (2) have a minimum of 100 approved HITs, (3) be located in Australia, Canada, India, the United Kingdom or the United States. The second prerequisite is to ensure workers have some experience in the completion of HITs. The first condition is to exclude those who often perform poorly and fail to meet the requester's demand, resulting in rejection of work. We select five classic countries that have a large number of English speakers. Meanwhile, it also provides us the possibility to explore the differences in language expression of different countries. So, for workers interested in our HIT, they have to fulfill all the above qualification requirements to accept the HIT.

There are four topic-oriented search scenarios in the study, namely argument, news, book and trip. They are quite common search topics in real life. We initially plan to recruit a total of 400 workers, 20 workers for each scenario (a total of 4) for each country (a total of 5). In the pilot study alpha, the estimated time for completing our assignment is around 15 minutes. So we set the time that per worker can work on is 1 hour to leave them more flexibility. According to the minimum wages (US$) by five countries recorded in wikipedia[1],

[1] https://en.wikipedia.org/wiki/List_of_minimum_wages_by_country

the rewards per assignment for five countries Australia, Canada, India, the United Kingdom and the United States are set to 3.64$, 2.18$, 0.25$, 2.74$ and 1.8$ respectively. Together with the additional payment to Amazon, our research budget is 1200$. However, since it is slower to get sufficient qualified workers for Australia than the other four countries in the formal news study, to motivate workers, the reward for Australia is increased from 3.64$ to 4.0$ per assignment.

Generally, we create a project called 'Conversational Queries' in the Amazon Mturk Requester interface and enter the predetermined properties such as the worker requirements, number of respondents, and reward per response, as mentioned above. A layout that contains the interface and script of the study is required to design and preview for confirmation. The overall setting of properties can be in the following:

- Title: Talk to a Search Bot
- Description: In this 10 minute task, communicate with the chatbot system using the text field and the green button. The task changes whenever you send a message. You are only allowed to do this task once.
- Keywords: user study, chat, search
- Reward per response: $2.74
- Number of respondents: 20
- Time allotted per Worker: 1 Hour
- Survey expires in 3 Days.
- Auto-approve and pay Workers in 7 Days.
- HIT Approval Rate (%) for all Requesters' HITs greater than 95.
- Number of HITs Approved greater than 100.
- Location is UNITED KINGDOM (GB).
- Task Visibility: Public. All Workers can see and preview my tasks.

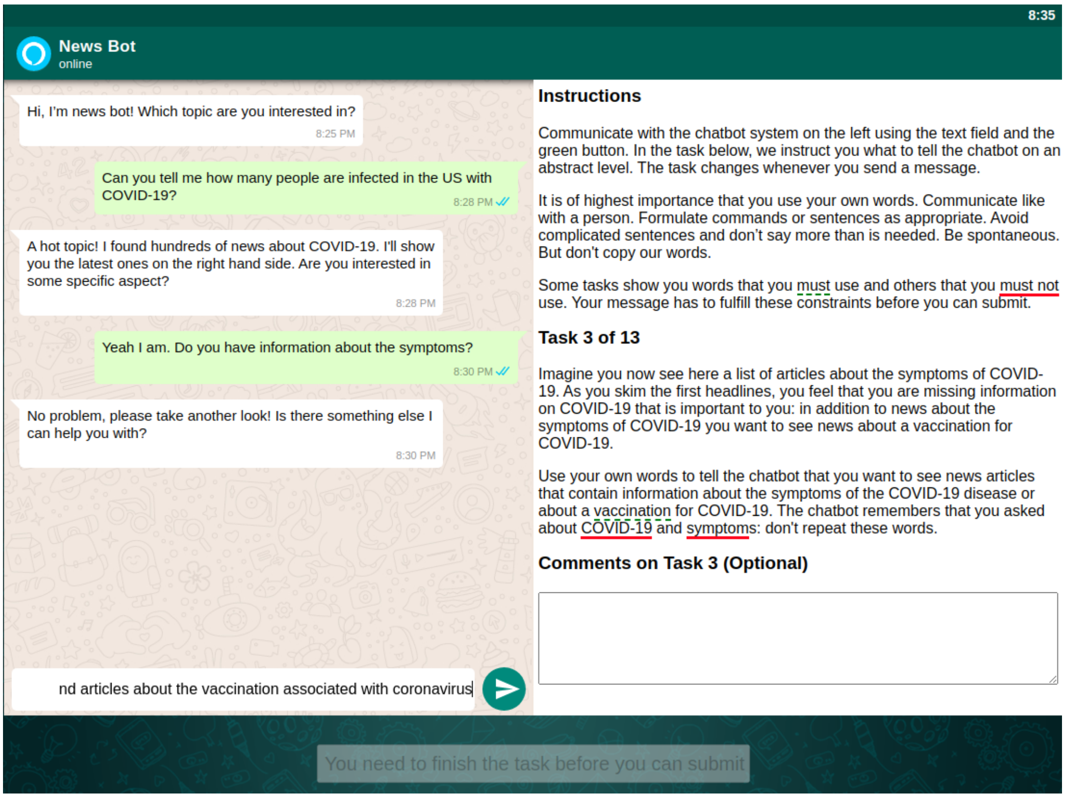
It is unnecessary to create a project for each scenario or for each country as we can publish multiple batches for a project. Before publishing a batch, we modify the reward and location settings correspondingly, the number of respondents as optional, and keep other properties unchanged. Every worker is only allowed to participate in the study with one of the scenarios to ensure the diversity of utterances, so we have to manually add their unique WorkerIDs in the script's block function while creating batch for the same scenario or for the other three scenarios. As a result, they can not view the study even if they accept the assignment in different batches. In the same batch, participants are only able to accept the assignment once. In the end, we can publish a batch by uploading a CSV file containing all necessary data of the study with one of the scenarios.

4.1.2 Pilot Study

The study focuses on how seekers perform query reformulation for various intents in topic-oriented search scenarios, but no real search is involved. We decide to design an interface that imitates human and human interaction like the messaging platform "Whatsapp". The participants are informed that they are communicating with a chatbot and need to perform a sequence of predefined tasks. After sending a text query for a task, the participants are told that they are given some results, but they are faced with a different situation, so they need to modify the query in a certain way.

We advance the interface and task descriptions by carrying out a total of 8 pilot studies with the news scenario and evaluating participants' performance. In each pilot study, we only choose five workers from India and the United States as they are two of the five countries with relatively row rewards. Every worker is only allowed to conduct one of the pilot studies. In the formal study, all workers are welcome to participate in the study again, even if they have done the pilot study. Once the interface and the task descriptions are confirmed for the formal study, we modify specific keywords used in the task descriptions for the other three scenarios.

In pilot study alpha, we design a chat-interface (Figure 4.1) consist of two separate parts. The left part presents the dialogue between participant and chatbot. The upper right part shows the instructions of the entire study and remains consistent in subsequent interactions. Participants type a query in the text field for a task and press the green button to send. Then the bottom right part goes to the next task and changes the index and description of the task correspondingly. Meanwhile, participants receive a predetermined response from the chatbot to remind them to view the new task on the right-hand side. They can write comments either for every single task or for the overall study after completing all tasks, and in the end, click the button at



Hi, I'm news bot! Which topic are you interested in?

Can you tell me how many people are infected in the US with COVID-19?

A hot topic! I found hundreds of news about COVID-19. I'll show you the latest ones on the right hand side. Are you interested in some specific aspect?

Yeah I am. Do you have information about the symptoms?

No problem, please take another look! Is there something else I can help you with?

nd articles about the vaccination associated with coronavirus

Instructions

Communicate with the chatbot system on the left using the text field and the green button. In the task below, we instruct you what to tell the chatbot on an abstract level. The task changes whenever you send a message.

It is of highest importance that you use your own words. Communicate like with a person. Formulate commands or sentences as appropriate. Avoid complicated sentences and don't say more than is needed. Be spontaneous. But don't copy our words.

Some tasks show you words that you must use and others that you must not use. Your message has to fulfill these constraints before you can submit.

Task 3 of 13

Imagine you now see here a list of articles about the symptoms of COVID-19. As you skim the first headlines, you feel that you are missing information on COVID-19 that is important to you: in addition to news about the symptoms of COVID-19 you want to see news about a vaccination for COVID-19.

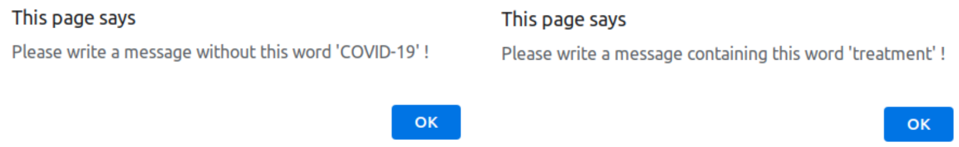
Use your own words to tell the chatbot that you want to see news articles that contain information about the symptoms of the COVID-19 disease or about a vaccination for COVID-19. The chatbot remembers that you asked about COVID-19 and symptoms: don't repeat these words.

Comments on Task 3 (Optional)

Figure 4.1: Designed interface in pilot study alpha.

the bottom to submit the study. There are some details in the interface, such as the real-time at the top right, the chatbot name, online status with head portrait at the top left, the time of received and sent messages, and the status of already read message with a tick. Interestingly, there is a loader that lasts a few seconds at the right of the input field after clicking the green button, which acts as the state of sending a message. All these details are aimed to make the participants more like being in a conversation.

Each task is comprised of three components: response from the chatbot, task scenario, and task prompt. As the example task is shown in Figure 4.1, in the task scenario, starting with '*Imagine you now see here a list of articles about...*', participants will not see the specific results but are asked to image they have seen a list. And they somehow changed their mind and had a new idea for the search. The task prompt, starting with '*Use your own words to tell the chatbot that you want to see...*' describes how they should reformulate their utterance to continue the search.


This page says

Please write a message without this word 'COVID-19' !

This page says

Please write a message containing this word 'treatment' !

(a) words with a red underline

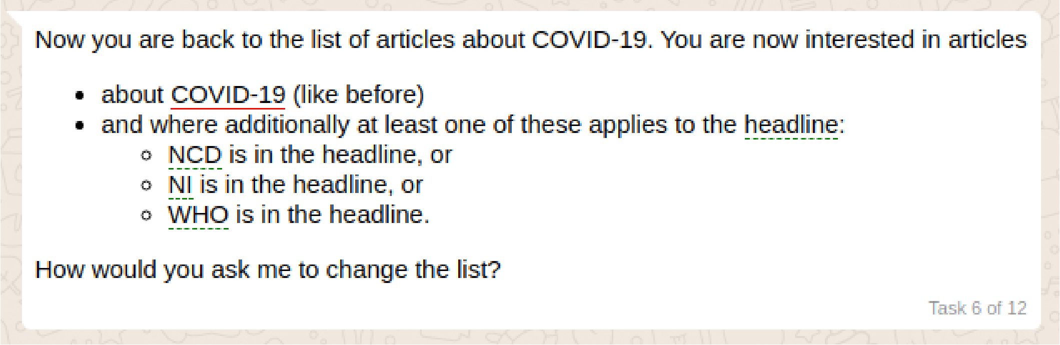
(b) words with a green dashed underline

Figure 4.2: Two kinds of popup messages to remind the participants about the use of keywords.

In the description of the task, there are two kinds of highlighting: words with a green dashed underline and words with a red underline. Red underlines denote the keywords which should not be used for the current task as they are already specified in previous turns. Green dashed underlines denote the keywords of the current task, so participants must include them in their utterances. The meanings of these two kinds of highlighting are also explained in the instructions. If participants do not follow the rules, there are corresponding popup messages to remind them (as shown in Figure 4.2). We would like to observe how seekers reformulate their utterance without retelling the already mentioned concepts (since the system should already ’know’ them), so it is helpful to emphasize these keywords to participants. Moreover, in this way, it is easier for us to identify participants who try to cheat to get the rewards because they only type the keywords as the answers to prevent the popup messages. So we keep the highlighting in further pilot studies and formal study as well.

Nevertheless, in the results of pilot study alpha, participants lose focus due to two separate parts of the interface since they are not used to switching attention on both sides. They have to first read the task on the right-hand side, type the utterance in the input field on the bottom of the left-hand side, read the above response from the chatbot and read the new task on the right-hand side again. Switching attention in such a way and also reading long text task descriptions might make them easy to lose patience and interest. Besides, Task 11, which asks participants to confirm their query, is removed after pilot study alpha since there is no diversity in the utterances.

In the pilot study beta, the interface with two separate parts is discarded. Instead, the tasks and instruction are embedded in the dialogue between participants and the chatbot. The instruction is at the beginning of the dialogue, and participants are told in the next bubble: ’*If you understood, tell me that you are ready*’ as the first task. They have to enter an utterance containing



Now you are back to the list of articles about COVID-19. You are now interested in articles

- about COVID-19 (like before)
- and where additionally at least one of these applies to the headline:
    - NCD is in the headline, or
    - NI is in the headline, or
    - WHO is in the headline.

How would you ask me to change the list?

Figure 4.3: Divide requirements of the task into sub-items in a list.

'*ready*' to move to the next task. The descriptions of the tasks are shortened by removing redundant task scenarios. They are only notified that the chatbot has changed the list accordingly, followed by the requirement of a new task and a question '*How would you ask me to change the list?*' in the end. The text field for comment of each task in the pilot study alpha is removed, but participants are guided to write down a comment after finishing all tasks by '*Thank you for your participation in this study! We would appreciate it if you leave us some feedback below. Then submit with the usual button.*'. In such an interface, they are able to review the entire interaction process and specific tasks at any time while doing the study.

For those tasks that contain multiple requirements, the requirements are divided into separate sub-items in a list instead of in closely consecutive sentences for Task 4 and Task 6. A visual representation of the example task is shown in Figure 4.3. It can be seen from the results that it is clear and intuitive for participants to grasp the key points of the task. So this strategy is also applied to other tasks such as Task 5 and Task 11 in the pilot study gamma. The restriction on keyword '*symptoms*' in Task 4 is removed, since it is also interesting to see if they prefer mentioning both filter conditions at the same time. Task 4 is denoted as Task 3 in the pilot study alpha (Figure 4.1) due to the additional *ready* task in latter pilot studies.

In the subsequent pilot studies, our main contribution is to fine-tune the task descriptions and find optimal descriptions that indicate all fundamental requirements but are concise and easy to understand, especially for those tasks which participants often misunderstand. While describing the tasks, we are careful about the ways of expression by selecting uncommon but still understandable ones like '*where additionally at least one of these (keywords) applies to the headline...*' to avoid participants copying our words as much as possible.

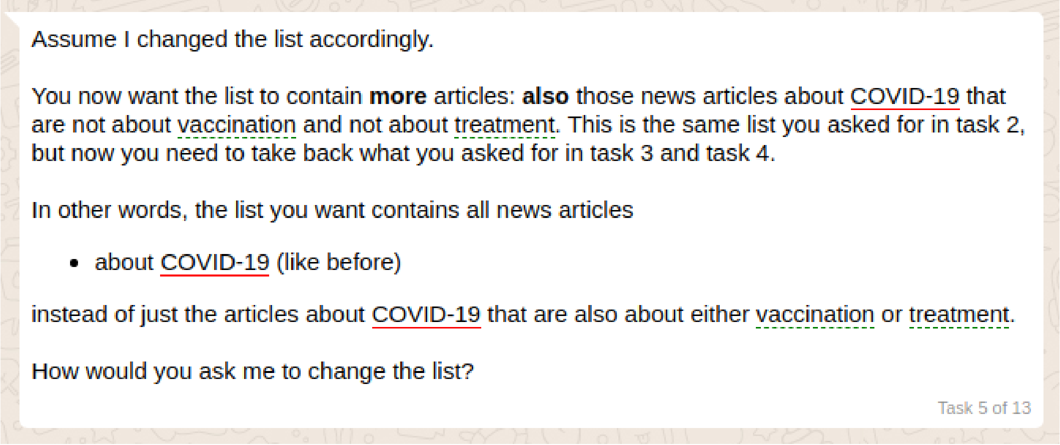
Meanwhile, it is crucial to have participants who have good English skills to ensure they can understand the task in a short time and present qualified answers. This requirement is clearly pointed out in the instruction as: '*You must have good English skills to work on this HIT: If we can not understand your messages or they are not what we asked for, we can not accept your work.*', which provides us the reason to reject their works in the further curation.

In the pilot study delta, the keyword in Task 3 is replaced from *symptoms* to *vaccination* and the keyword in Task 4 is replaced from *vaccination* to *treatment* as seekers are more likely to ask for news about vaccination and treatment of COVID-19 than the news about symptoms and vaccination of COVID-19. This makes the context of the tasks more reasonable.

For Task 11, participants are required to replace the keyword conditions for the headline that are specified a few turns ago (Task 6) with another new filter condition *not economy*. In other words, It is no longer necessary for the chatbot to specify the headline of news with certain keywords but to filter out all news containing the keyword *economy*. However, it seems that the requirements are too complicated to understand for the participants. Some of them forget to mention the negation of the condition *economy* or misunderstand the keyword *condition* should not be in the headline. After several attempts, only a few participants give the correct utterances. In the pilot study epsilon, we remove the negation of the condition *economy* and add a new Task 13 that is to ask for news about flu (Task 12) but not about COVID-19, which represents the operation Create Literal accordingly. For tasks that include negation like Task 7 and Task 13, **not** is highlighted as bold.

For Task 5, there are similar misunderstandings problems. Participants misunderstand the task as specifying news without keyword *vaccination* or *treatment* instead of removing these two keywords from filters. That is to say, the list should include all news about COVID-19 again, whether it is about *vaccination* or *treatment*. In the pilot study zeta, we try to bring the task scenario back to the task description and also add some determiners such as **more**, **less** or **different** to describe the returned list (Figure 4.4). After the pilot study zeta and eta, it can be found that adding back the task scenario does not help participants better understand the task, but it is more likely to increase the complexity of the task. We remove the task scenario but keep the determiners in the final study.

Another task that participants perform suboptimally is Task 10. Participants are informed that they forgot their instructions for the search results



Assume I changed the list accordingly.

You now want the list to contain **more** articles: **also** those news articles about COVID-19 that are not about vaccination and not about treatment. This is the same list you asked for in task 2, but now you need to take back what you asked for in task 3 and task 4.

In other words, the list you want contains all news articles

- about COVID-19 (like before)

instead of just the articles about COVID-19 that are also about either vaccination or treatment.

How would you ask me to change the list?

Figure 4.4: Task description with the task scenario in pilot study zeta.

and '*How would you ask me to tell you your previous instructions?*'. They are expected to inquiry about their cumulative query or their filter conditions specified before. However, they either do not understand the task or give the utterances that are not what we expected. It is difficult to ensure that most people understand this task because 'query' and 'filter' are technical terms and are probably unknown for people who do not have much scientific background or have little knowledge about information retrieval.

In the instruction, in addition to introducing the meanings of two highlighting for words and required good English skills, the capability of the chatbot is supplemented. The chatbot can not only find news on the internet but also '*remember what you asked for*', which emphasizes to participants that they only need to tell the chatbot the changes from their last message. COVID-19 is spreading all over the world while we are doing the pilot study. Participants perform actively in the search and indeed have many individual ideas about the topic, notably in Task 2. They add various extra information in their utterances, for instance, *the death rate of COVID-19*, *the number of patients suffering from COVID-19 in India*, *the spread situation of COVID-19.* As these utterances are usually in the form of questions, they are more fit to another area of information-seeking: question answering (also refer to fact finding [Kellar et al., 2007]) instead of asking for returning a collection about a certain topic in our case (information gathering). In the pilot study zeta, an independent paragraph is given in the instruction, which emphasizes '*Do not add unnecessary information to your messages.*' and provides some specific unnecessary information that added in Task 2 as examples.

4.1.3 Formal Study

The final study combines the parts from the pilot studies that participants perform well and keeps all task descriptions consistent. The complete instructions for the news scenario are in the following:

Hi! Imagine that I'm a chatbot that you can ask to find you news articles on the Internet. However, in this study, I actually won't show you the news articles to keep the study short. In this study, I will tell you to ask me for a specific list of news articles. And then, I will tell you to ask me to add or remove (or both) certain news articles from the list. Unlike common search engines (e.g., Google), I remember what you asked for: so you just need to tell me how the list should change. In order to make clear what needs to change, I will underline it like this: you **must** use these words in your message. And I will underline like this what must stay the same: you **must not** use these words **nor synonyms** of them in your new message. Don't add unnecessary information to your messages. For example, you will have to ask me for a list of news articles about a disease. Do not ask for news articles 'from 2020' or 'about infections' or 'the death rate' or other more specific information about the disease. Just ask for news articles about the disease. But otherwise, formulate your own messages as you would for a real chatbot! You must have good English skills to work on this HIT: If we can not understand your messages or they are not what we asked for, we can not accept your work.

For other scenarios, the collection type news articles in the instructions are replaced by arguments, books or trips.

Table 4.1 shows the detailed description and the expected operation of each task in the news scenario.

Task	Description	Expected Operation
1	If you understood, tell me that you are ready.	$Q_0 : \perp$
2	Good! How would you ask me to get you all news articles on the COVID-19 disease?	Create(l_1)

3	Imagine I would show you the list of all news articles about COVID-19 as you asked for. You now want the list to contain **fewer** articles: It should contain just the news articles about COVID-19 (you already told me) that are about a vaccination. How would you ask me to change the list?	Create(l_2)
4	Assume I changed the list accordingly. You now want the list to contain **more** articles: It should contain the news articles about COVID-19 (you already told me) that are about either: vaccination (you already told me), or treatment (as a new alternative to vaccination). How would you ask me to change the list?	Update(l_2, $l_2 \vee l_3$)
5	Assume I changed the list accordingly. You now want the list to contain **more** articles: It should again contain all news articles about COVID-19 (you already told me, but now again with those news articles that are not about vaccination or treatment). How would you ask me to change the list?	Delete($l_2 \vee l_3$)
6	Now you are back to the list of articles about COVID-19. You now want the list to contain **fewer** articles: It should contain just the news articles about COVID-19 (you already told me) for which also at least one of these applies to the headline: NCD is in the headline, or NI is in the headline, or WHO is in the headline. How would you ask me to change the list?	Create(f_1:($l_4 \vee l_5 \vee l_6$))
7	Assume I changed the list but you see that I understood 'NI' as meaning 'North Ireland', but this is not what you had in mind. How would you tell me? Do **not** tell me yet what you had in mind (in case you have an idea).	Update(l_5, ?)
8	Assume you had the National Insurance in mind. How would you answer me if I now asked: 'What did you actually mean?'	Update(?, l_7)

9	Assume I changed the list accordingly. You now want the list to contain **different** articles: It should contain the news articles about SARS-CoV-2 (instead of COVID-19) for which also at least one of these applies to the headline: (you already told me) NCD is in the headline, or NI (meaning 'National Insurance') is in the headline, or WHO is in the headline. How would you ask me to change the list?	Update(l_1, l_8)
10	Assume I changed the list accordingly and that you read some news articles. You then come back to me and forgot your instructions (that the list should contain the articles about SARS-CoV-2 where the headline must contain at least one of NCD, NI, or WHO). How would you ask me to tell you your previous instructions?	Read(Q_{10})
11	You now want the list to contain **different** articles: It should contain just the news articles about SARS-CoV-2 (you already told me) that are about the economy (instead of having a headline that contains at least one of NCD, NI, or WHO.) How would you ask me to change the list?	Update(f_1:($l_4 \vee l_7 \vee l_6$), l_9)
12	Assume I changed the list accordingly. You now want the list to contain **different** articles: It should contain the news articles about flu (I should not consider anything you said earlier). How would you ask me to change the list?	Update(Q_{11}, l_{10})
13	Assume I changed the list accordingly. Finally you want the list to contain **fewer** articles: It should contain just the news articles about the flu (you already told me) that are **not** about COVID-19. How would you ask me to change the list?	Create($\neg l_1$)

Table 4.1: Task descriptions and expected operations performed by participants in the news-oriented search.

For the other three scenarios, we keep the structure of task descriptions consistent but modify the keywords used in the task descriptions. Table 4.2 shows the specific keywords (literals) used in different scenarios.

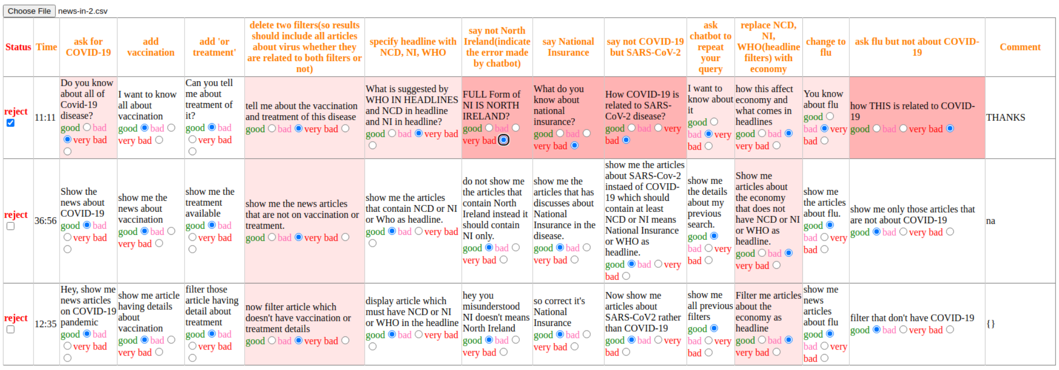
S	argument	book	news	trip
l_1	banning plastic bags	virus	COVID-19	San Jose
l_2	CO2 emissions	infected animals	vaccination	by ship
l_3	renewable resources	plants	treatment	by car
l_4	BestReasons	Thriller	NCD	Hilton
l_5	WhatsUp	SF	NI	HI
l_6	WikiDiscussions	Horror	WHO	BW
f_1	source	genre	headline	hotel
?	news page	San Francisco	North Ireland	Hampton Inn
l_7	discussion forum	Science Fiction	National Insurance	Holiday Inn
l_8	subsidizing paper bags	mutants	SARS-CoV-2	San Antonio
l_9	fashion	scientific background	economy	sightseeing
l_{10}	banning plastic drinking straws	evolution	flu	Santiago

Table 4.2: The specific values of Symbols (S) in four search scenarios including a series of literals l_i, misrecognized word '?', field name defined in the function f_1.

4.2 Curation of Conversational Query Reformulation

For reviewing the works done by participants, we designed a specific interface. The results can be downloaded from the Manage Batches interface as a CSV file, which is uploaded to the curation interface to create a form containing all data (Figure 4.5). Each row presents information of a participant, including the time they work on the study, answer for each task and comment. The prompt of each task is shown in the header.

The utterances are classified into three categories: "good", "bad", or "very bad". "Very bad" utterances will be discarded in the database, even if they are from a worker who is approved. A "very bad" utterance can be an answer consisting of only keywords that are underlined in red in the task description, as it seems like the participant would like to cheat. Some of the participants might paste a small paragraph from the Internet that includes the required keywords as the answer, but the content fails to meet the task's requirement and context. These answers that do not make sense themselves or are not related to the tasks will be regarded as "very bad". To be more specific, for instance, in Task 2 of the study with the news scenario, which asks for a list of news, the utterance *Please find me the total numbers of COVID-19 today* is "very bad", as it does not mention news at all and contains additional information as well. In Task 7, it has been specified that North Ireland is not what you had in mind, if the utterance is *NI is North Ireland*, which do not indicate the negation of North Ireland, such an utterance is "very bad". Similarly, in



Choose File news-in-2.csv

Status	Time	ask for COVID-19	add vaccination	add 'or treatment'	delete two filters(so results should include all articles about virus whether they are related to both filters or not)	specify headline with NCD, NI, WHO	say not North Ireland(indicate the error made by chatbot)	say National Insurance	say not COVID-19 but SARS-CoV-2	ask chatbot to repeat your query	replace NCD, NI, WHO(headline filters) with economy	change to flu	ask flu but not about COVID-19	Comment
reject ☑	11:11	Do you know about all of Covid-19 disease? good ○ bad ● very bad ○	I want to know all about vaccination good ● bad ○ very bad ○	Can you tell me about treatment of it? good ● bad ○ very bad ○	tell me about the vaccination and treatment of this disease good ○ bad ● very bad ○	What is suggested by WHO IN HEADLINES and NCD in headline and NI in headline? good ○ bad ● very bad ○	FULL Form of NI IS NORTH IRELAND? good ○ bad ○ very bad ●	What do you know about national insurance? good ○ bad ○ very bad ●	How COVID-19 is related to SARS-CoV-2 disease? good ○ bad ○ very bad ●	I want to know about it good ○ bad ● very bad ○	how this affect economy and what comes in headlines good ○ bad ● very bad ○	You know about flu good ○ bad ● very bad ○	how THIS is related to COVID-19 good ○ bad ○ very bad ●	THANKS
reject ☐	36:56	Show the news about COVID-19 good ● bad ○ very bad ○	show me the news about vaccination good ● bad ○ very bad ○	show me the treatment available good ● bad ○ very bad ○	show me the news articles that are not on vaccination or treatment. good ○ bad ● very bad ○	show me the articles that contain NCD or NI or Who as headline. good ● bad ○ very bad ○	do not show me the articles that contain North Ireland instead it should contain NI only. good ● bad ○ very bad ○	show me the articles that has discusses about National Insurance in the disease. good ● bad ○ very bad ○	show me the articles about SARS-Cov-2 instaed of COVID-19 which should contain at least NCD or NI means National Insurance or WHO as headline. good ● bad ○ very bad ○	show me the details about my previous search. good ● bad ○ very bad ○	Show me articles about the economy that does not have NCD or NI or WHO as headline. good ○ bad ● very bad ○	show me the articles about flu. good ● bad ○ very bad ○	show me only those articles that are not about COVID-19 good ● bad ○ very bad ○	na
reject ☐	12:35	Hey, show me news articles on COVID-19 pandemic good ● bad ○ very bad ○	show me article having details about vaccination good ● bad ○ very bad ○	filter those article having detail about treatment good ● bad ○ very bad ○	now filter article which doesn't have vaccination or treatment details good ○ bad ● very bad ○	display article which must have NCD or NI or WHO in the headline good ● bad ○ very bad ○	hey you misunderstood NI doesn't means North Ireland good ● bad ○ very bad ○	so correct it's National Insurance good ● bad ○ very bad ○	Now show me articles about SARS-CoV2 rather than COVID-19 good ● bad ○ very bad ○	show me all previous filters good ● bad ○ very bad ○	Filter me articles about the economy as headline good ○ bad ● very bad ○	show me news articles about flu good ● bad ○ very bad ○	filter that don't have COVID-19 good ● bad ○ very bad ○	{}

Figure 4.5: Interface for reviewing utterances of participants.

Task 13, participants are expected to say not about COVID-19. Utterances like *List about COVID-19* are "very bad" as the negation "**not**" is emphasized as bold in the task description. For Task 9, one of the requirements is news should be about SARS-CoV-2 instead of COVID-19. These utterances such as *show news articles related to SARS-CoV-2 and COVID-19* or *SARS-CoV-2 is also dangerous virus like COVID-19* are considered as "very bad", because they are far from the requirement of the task. For the above relatively easy and unambiguous tasks, we have ensured that most participants can understand the task descriptions correctly in the pilot studies, so participants in the formal study should also at least give the correct answers for these tasks. However, we are tolerant of some tasks that might be difficult for participants to understand like Task 5 (delete filters), Task 10 (read query), and Task 11 (replace filters). As long as the utterances are relevant to the context of the tasks, they will be judged as "good" or "bad".

A general criterion for "good" or "bad" utterance is the fulfillment of the task requirements. For some tricky tasks, even if the utterances do not satisfy the task requirements, as long as the participants have tried to complete the tasks, these utterances are regarded as "bad" rather than "very bad". If an utterance incorporates additional information beyond the task's requirements and it still makes sense after removing the unnecessary information, it will be regarded as "good", otherwise "bad". For example, in Task 3 of the news scenario, which requires to filter the result list with vaccination, some "bad" utterances are *please remove articles about vaccination* and *Is there a formulated vaccination for the virus?*. The second example does not meet the task requirement after eliminating unnecessary information (*formulated*, *virus*). Other "bad" utterances in the argument scenario can be *add arguments about CO2 emissions to the list* and *can you also tell me arguments that are about CO2 emissions?*, as they obviously refer to the operation Update(l_1, $l_1 \lor l_2$) instead of Create(l_2). In other words, the list includes arguments about banning plastic bags or arguments about CO2 emissions. However, as expected, CO2 emissions should be a subtopic based upon the main-topic banning plastic bags. An utterance for Task 4 like *or display news articles on treatment and progress* is "good" after removing the additional information *and progress*. In Task 4, participants are free to mention both l_2 and l_3. But if the participants say *I need the list to contain books about infected animals and plants* or *get books that are also about plants* in the book scenario, the utterances are regarded as "bad", as they are interpreted as the operation Update(l_2, $l_2 \land l_3$) rather than Update(l_2, $l_2 \lor l_3$). The studies do not allow participants to repeat concepts that are already mentioned in previous turns. But some participants try to use alternative words to replace these concepts. The replacement of *virus* or *corona* with *COVID-19* is not acceptable, while the usage of pronoun is allowed such as *show more details about its vaccination* in Task 3.

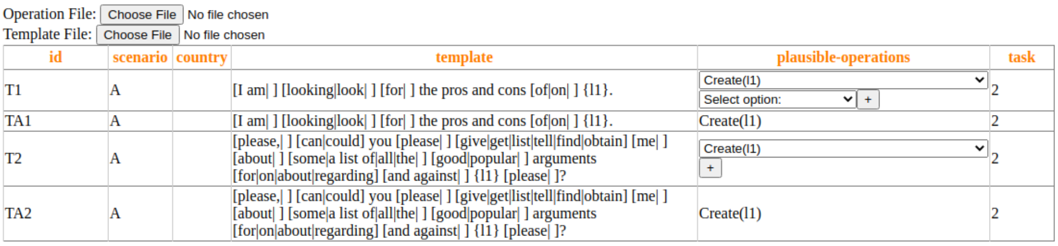
In the description of Task 9, participants are required to replace literal l_1 with another new literal l_8 and are free to mention the conditions specified in previous turns. What is more concerned is how they use alternative expressions to indicate the already mentioned conditions rather than repeating them exactly (e.g., *with the same headlines as before* instead of *with NCD, NI, WHO in the headline*). In Task 10, some nouns referring to the search history are expected like *search criteria, last query, request parameters, commands.* The utterances that ask for repeating the result list are "bad" as *return to my previous search results.* In Task 12, participants need to start a new search for another main-topic, but it does not matter whether they emphasize that it should be a new search or all previous search parameters are removed. In Task 5 (remove filters) and Task 11 (replace filters) for all scenarios, there are a large number of misunderstanding utterances, which are considered as "bad". However, these "bad" utterances are sorted into two additional intents. For Task 2 in the argument scenario, utterances that prefer towards one aspect of arguments are not expected like *what are the reasons pro banning plastic bags?*, since it has been pointed out that '*find you arguments (pros and cons)*' in the instruction.

These three kinds of utterances ("good", "bad" and "very bad") are distinguished by color in the curation interface. If an utterance is selected as "bad", the background color of this text area is filled with shallow red. If the utterance is selected as "very bad", the whole text area is colored as dark red. "Good" utterance has no specific background color. If a participant has over three "very bad" answers, a reject checkbox in the first column can be selected. At the end of the interface, there are some buttons in the interface for different needs such as exporting all "good" and "bad" utterances as a CSV file, exporting "good" utterances as a CSV file, exporting "bad" utterances as a CSV file, downloading the interface including all curation results, which is convenient to modify some of the results when needed without restarting review for all utterances, and a button to generate the rejection or approval status of all participants as a CSV file. For a participant who is rejected, there is a rejection message to explain why he or she is rejected. The structure of a rejection message is as the following:

"We had a look at your responses. Unfortunately, you did not do what we asked for several times. For example, we asked you to" + **prompt** + "but you said" + **utterance**.

They will not be told all tasks they perform poorly but only one of the tasks as an example. Our goal is to reject participants who attempt to cheat to get rewards or lack good English skills and to approve participants who accomplish the study well or try their best. The answers from the rejected workers will not be taken into account in the database. Participants, who receive a rejection message, are allowed to send a request for asking the requester to review their work again by mail. After the review, if they indeed do not satisfy the study's requirements, we will give them another task that they perform badly as examples to convince them. Participants should not only include all required keywords to prevent the popup messages but also fully understand the task and fulfill the requirements in the answers.

Although we indicate the good English skill as one of the necessary requirements for completing the study in the instruction, the results for India are not satisfactory. In the study with the news scenario, there are a total of 3 batches creating for India. Unfortunately, the rejection rate of each batch is higher than expected. There are 40 of 47 participants rejected in batch news-in-1, with a rejection rate of 85.11%. The rejection rate of batch news-in-2 is as high as 76.92% and of batch news-in-3 is 65.63%. For the study with the argument scenario, in the batch argument-in-1, the rejection rate is still high. Meanwhile, we received many emails from the workers in India complaining about their works are being rejected. In order not to upset more Indian workers, we decide to stop the study for India in the study with arguments and will not take India into account in the last two search scenarios (book, trip). In addition, it takes almost two months (from 18th June to 2nd Aug.) to get sufficient participants for Australia in the study with the news scenario. We thus canceled the studies for the other three scenarios for Australia.



Operation File: Choose File No file chosen
Template File: Choose File No file chosen

<table>
<tr><th>id</th><th>scenario</th><th>country</th><th>template</th><th>plausible-operations</th><th>task</th></tr>
<tr><td>T1</td><td>A</td><td></td><td>[I am|] [looking|look|] [for|] the pros and cons [of|on|] {l1}.</td><td>Create(l1)
Select option: +</td><td>2</td></tr>
<tr><td>TA1</td><td>A</td><td></td><td>[I am|] [looking|look|] [for|] the pros and cons [of|on|] {l1}.</td><td>Create(l1)</td><td>2</td></tr>
<tr><td>T2</td><td>A</td><td></td><td>[please,|] [can|could] you [please|] [give|get|list|tell|find|obtain] [me|] [about|] [some|a list of|all|the|] [good|popular|] arguments [for|on|about|regarding] [and against|] {l1} [please|]?</td><td>Create(l1)
+</td><td>2</td></tr>
<tr><td>TA2</td><td>A</td><td></td><td>[please,|] [can|could] you [please|] [give|get|list|tell|find|obtain] [me|] [about|] [some|a list of|all|the|] [good|popular|] arguments [for|on|about|regarding] [and against|] {l1} [please|]?</td><td>Create(l1)</td><td>2</td></tr>
</table>

Figure 4.6: Interface for reviewing the utterances of the participants.

After the curation for all utterances, our contribution is to organize the "good" utterances as various templates according to their intents. Several utterances are summarized into a template due to the similar sentence structure. For example, the same subject 'I', same sentence pattern 'can I' or similar verbs like 'get', 'show'. However, if the sentence is too long or has its unique structure, it will be regarded as an independent template. In the news scenario, the templates are distinguished by five countries. In the other three scenarios (argument, book, and trip), the templates from 3 countries (Canada, the United Kingdom, and the United States) are merged together. A dataset for collecting all templates is generated. Each template has its unique ID, scenario and the task it belongs to. However, "good" utterances are likely to be ambiguous. In other words, they are interpretable and meet the requirements in this task but also can be used for other tasks with different intents and operations. They are not excluded from the "good" classification. Instead, a dataset is created for further analysis, collecting all "good" utterances with possible operations, their belonger (worker with a unique ID), the task they belong to, type of utterance, and their matching template IDs, corresponding to the templates dataset. In addition, the utterances will be modified for typos, and the unnecessary information is taken away from templates and utterances. An interface is designed to efficiently assign the operations for each template by choosing an operation from a select list. With clicking a plus button, a new select list of possible operations is available (Figure 4.6). Another similar interface is designed for assigning the templates for each utterance by select list and choosing the type of utterance from **question**, **command**, **statement**. The possible operations will be automatically generated after selecting the templates. The results for the dataset are downloaded as CSV files, and all operations performed in the interface can be saved locally for further quick modifications.

Overall, in the study with four scenarios for five countries, there are a total of 284 approved workers. 4 of them are out of the plan after reviewing works due to the overdemanding criteria in the early formal study with the news scenario. There are 2919 utterances that are labeled as "good", and 1434 templates are organized. The analysis of the collected reformulation patterns will be elaborated on in Chapter 6.

Chapter 5

Prototyping the Conversational Query Rewriting Layer

To implement the theoretical model Query Rewriting Layer proposed in Section 3.2 that enables seekers to pose complex queries in conversational search, we introduce a prototype developed with Alexa Skills Kit and search Engine Elasticsearch in this chapter. Using Alexa Skills Kit, a skill with a custom interaction model is created in Section 5.1, which absorbs the utterances that are collected in the study (Chapter 4). We present the idea of how the system takes advantage of Elasticsearch in the back-end to rewrite the existing queries recursively while seekers are interacting with the system in Section 5.2. In Section 5.3, we perform a cross-evaluation to explore the generalizability of the reformulation patterns collected in different scenarios in Alexa Developer Console.

5.1 The Conversational Query Interaction Model

To create an interaction model for a custom skill, we requires to declare **Intents**, **Sample utterances**, **Custom slot Types** and **Dialog model** as optional[1]. **Intents:** represent actions that users can execute with the skill. **Sample utterances:** specify a series of words and phrases users can say to trigger the intents. **Custom slot Types:** a representative list of possible values for a slot that can be embedded in sample utterances. **Dialog model:** a structure that identifies information the skill requires and the prompts Alexa can use to collect and confirm that information in a conversation with the user.

[1] https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-the-interaction-model-for-your-skill.html

In the study (Chapter 4), we collect utterances for 12 reformulation tasks in four topic-oriented search scenarios and remove those are ambiguous, misunderstood by participants, or do not satisfy the requirements of each task. The remaining utterances are summarized as patterns according to the similar syntax. The skill consists of 11 custom Intents, which corresponds to each task except Task 8. An overview of all intents and their corresponding tasks and properties can be seen in Table 5.1. The collected reformulation patterns of each task are adopted as sample utterances in the corresponding intents. A pattern in the sample utterances of CreateQueryIntent (Task 2) is [please|] [show|get|gather|give] [me|] [all|a list of|] [recent|latest|] {collection} [about|on] {filter}, where {collection} and {filter} are two custom slot types. [about|on] represents it can be either about or on in this position and [please|] represents it can be either please or nothing here. {collection} includes all possible values of collection types that consistent with the scenarios as [arguments|books|news|trips] and {filter} includes values of all predetermined filters in tasks shown in Table 4.2. Unfortunately, the size of the generated interaction model file is five times as large as what Amazon allows such that it fails to build the interaction model in the Alexa Developer Console, because Alexa will generate enormous sample utterances to include all possibilities of patterns. Hence, we have to declare the collected utterances rather than the summarized patterns as sample utterances of intents at the expense of diversity of sample utterances.

Task	Intent	Utterances	Slots	Descriptions
2	CreateQueryIntent	274	2	Seekers start a search session for collection about a main-topic.
3	CreateLiteralIntent	82	2	Seekers add a subtopic to filter previously obtained result list.
4	UpdatePartIntent	55	3	Seekers add an alternative subtopic.
5	DeletePartIntent	70	3	Seekers remove part of filters.
6	CreatePartIntent	262	5	Seekers specify certain field with part of filters.
7	RejectLiteralIntent	155	4	Seekers indicate a error made by the system and reject the unexpected filter.
9	UpdateLiteralIntent	258	4	Seekers replace a existing filter with a new filter.
10	ReadQueryIntent	189	1	Seekers ask for system to recall the search history.
11	UpdatePartFieldIntent	133	6	Seekers replace part of filters for certain field with another new filter.
12	UpdateQueryIntent	140	2	Seekers replace the entire query with a new filter.
13	CreateNegLiteralIntent	263	2	Seekers add a negation of existing filter.

Table 5.1: Overview of all custom intents in the prototype and their corresponding tasks in the study, where task 8 is regarded as the confirmation part of RejectLiteralIntent (Task 7).

While designing the interaction model, there are other restrictions besides the limited file size of interaction model: (1) It is not allowed to exist any punctuations in the sample utterances of intents since sample utterances serve as users' spoken requests. The sample utterances we declared are collected in a text-based interface so that punctuations are inevitable. In particular, for those tasks with multiple requirements, participants prefer using multiple sentences with punctuations to convey the information needs. All punctuations in the utterances are replaced with empty space. However, eliminating punctuations can cause ambiguity. For instance, removing punctuations that connect multiple incomplete sentences like *no, the other NI, not North Ireland.* (2) Phrase slot type AMAZON.SearchQuery allows for input from a user with fewer constraints on format and content[2]. This is very suitable to be used to capture the less predictable input from a user. For example, in Task 8, seekers inform the system about the correct value for a certain filter specified before but misunderstood by the system. The correct value is hard to predict to build a list of possible values for the slot. Nevertheless, more than one phrase slot per intent is not allowed, and the phrase slot can not be embedded in a sample utterance with other custom slot types at the same time. Thus, alternatively, we only use custom slot types and all custom slot types in the interaction model are specified with the same list of possible values, including all collection types and all predetermined filters in the tasks (Table 4.2).

A name of an intent is designed based on its operation and its target presented in the Table 3.2. Although Task 4 and Task 11 have the same operation Update and target Part, Task 4 is to add an alternative filter l_3 in addition to the existing filter l_2 and Task 11 is to replace the whole part of the filters for a field with a new filter l_9. The intent for task 11 is renamed UpdatePartFieldIntent. Excluding the consideration of Task 8, Task 7 and Task 9 also have the same operation Update and target Literal. Task 7 is to reject the interpretation of certain filter given by the system, so we name the intent for Task 7 RejectFilterIntent. We add a negation in the name of Intent for Task 13 to distinguish from Task 3. After building the interaction model, it is possible to find the utterances conflicts in the model, representing that an utterance can be mapped into more than one intent. In other words, there are overlapping utterances in different intents such that Alexa is not able to distinguish these intents exactly. An example can be: `I do not want a {filter}` in RejectLiteralIntent when the system misinterprets a filter so that seekers do not want the unexpected value and the same utterances occurred in CreateNegLiteralIntent since seekers would like to add a restriction for the result list as a negation of a filter. CreateNegLiteralIntent aims to provide more details for the search, while RejectLiteralIntent provides negative feedback for the system's errors, and the

[2] https://developer.amazon.com/en-US/docs/alexa/custom-skills/slot-type-reference.html#phrase-types

system has to ask for confirmation about the required slot. The confirmation part of RejectLiteralIntent is equivalent to System Revealment proposed by [Radlinski and Craswell, 2017]. In previous turns, seekers generally take the initiative to instruct the system on modifying the existing query and indicating the system's error. However, the system takes the initiative, in turn, to ask for clarification about certain information to get more details. ReadQueryIntent shows a conversational search system's memory capability since it can refer to earlier statements and even navigate directive such as go back and repeat.

5.2 Prototype Back-End

A server serves as the back-end to mainly tackle with data storage and data searching using search engine Elasticsearch. With the help of RESTful API, all data or documents can be easily stored in JSON format and managed in the server. Elasticsearch correspondingly creates index for the imported data. Once the index is created, it is able to search for information according to indexing. Elaticsearch supports full text indexing and automatic detection of data structure and data types. Hence, the server not only eliminates restrictions on localized storage of documents but also facilitate the searching of data.

In a multi-turn conversational search session, how can the server handle the queries sent by seekers by referring to the context of the whole dialogue? It is well-known that natural language queries can be ambiguous. That is to say, there are multiple interpretations for a seeker query. For instance, an utterance *tell me articles about COVID-19 and vaccination or treatment* can be interpreted as asking for articles about both COVID-19 and vaccination or articles only about treatment as an alternative. Another explanation is that the returned list should contain articles only about COVID-19 as well as articles about either vaccination or treatment. It is also possible for seekers to clarify these complex information needs cumulatively with multiple turns. For example, in the first turn, they say *tell me articles about COVID-19.* Then in the second turn, an utterance can be *just tell me articles about its vaccination*, which contains coreference 'it' such that it is necessary to reference the context to figure out what 'it' refers to. Also, we classify the operations that can be performed to modify the previous queries into four types: Create, Read, Update, and Delete (known as acronym CRUD in Chapter 3). The corresponding intents of seekers could be adding more restrictions for the result list, navigating directives such as repeating earlier search details, changing minds to replace or releasing the old restrictions, or making the correction.

To solve the above problems, a Query class and its sub-classes are designed to provide the vocabulary used to formulate queries, which can be understood by search engine. In this way, natural languages utterances are transformed into queries of the computer languages in the search engine. There are several functions in the Query class like detecting ambiguity, minimizing the query for easier processing, and describing the query in natural language. As a result, the generated queries should be unambiguous and minimal. In a new search session, a simple query Q_1 transformed from an utterance *tell me articles about COVID-19* is in the following:

```
{
    "query": {
        "bool": {
            "must": [
                { "match": { "content": "COVID-19" }}
            ]
        }
    }
}
```

The "content" field of the documents (articles) must contain the word "COVID-19". Besides "must", other occurrence types in a boolean query are "filter", "should", and "must_not". "match" refers to perform a full-text search. The generated query string can be sent as a request through the REST client for Elasticsearch and receive a response containing a collection list. By default, Elasticsearch sorts the matching search results in the response by relevance score, which measures how well each document matches a query. In addition, it is possible to specify the number of matching search documents in the request. In each turn, several personal information of a seeker such as a unique user ID, device ID and session ID is stored in a UserRequest class. It will also save a query transformed from an utterance, the current query that might refer to the previous queries, the result list generated according to the current query, the utterance, and the triggered intent. All these required data are stored as the index, which is independent of another index for saving all collection documents that are returned as results. In other words, the system can keep track of previous queries and update the current query recursively during a conversational search session. After creating the query Q_1 mentioned in the above example, an utterance created in the next turn is *just tell me articles about its vaccination.* By reference to the previous query Q_1, the current query is as:

```
{
    "query": {
        "bool": {
            "must": [
                { "match": { "content": "COVID-19" }},
                { "match": { "content": "vaccination" }}
            ]
        }
    }
}
```

The "content" field of the documents must contain the word "COVID-19" as well as the word "vaccination". The recursive progress of updating the current query is shown in Figure 3.1. In such a way, the model Query Rewriting Layer (Section 3.2) is implemented to enable seekers to pose complex queries in a conversational search. One example of complex query is the current query of the sixth operation in Table 3.2, which represents as $l_1 \wedge f_1 : (l_4 \vee l_5 \vee l_6)$ using logical expression. f_1 specifies the field as headline rather than general content. The actual representation of the current query in the news scenario is:

```
{
    "query": {
        "bool": {
            "must": [
                {
                    "match": {"content": "COVID-19"}
                },
                {
                    "bool": {
                        "should": [
                            {"match": {"headline": "NCD"}},
                            {"match": {"headline": "NI"}},
                            {"match": {"headline": "WHO"}}
                        ]
                    }
                }
            ]
        }
    }
}
```

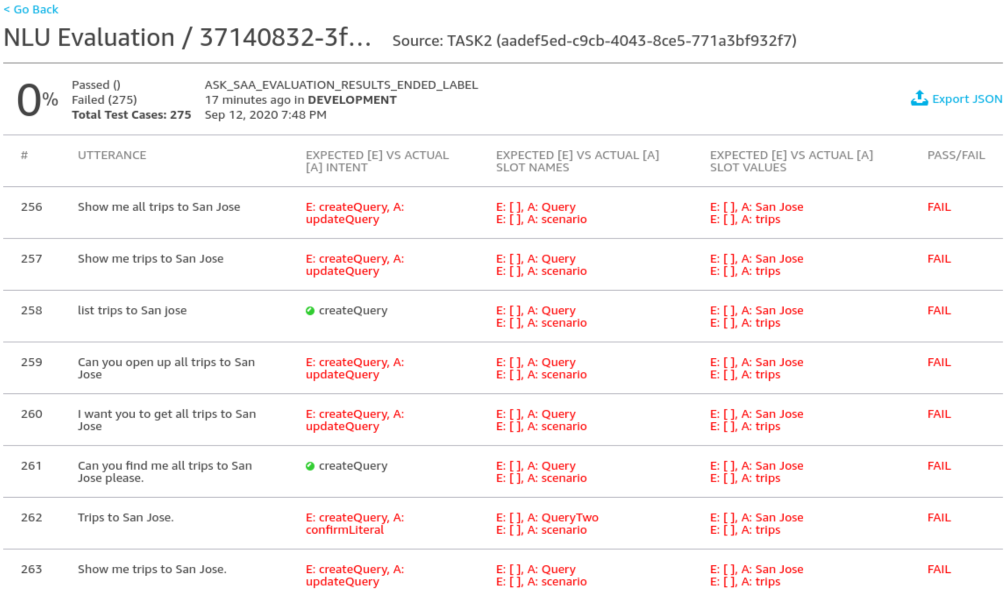
Although we did not implement the back-end of the prototype in detail yet, the general ideas presented in this section could be a promising directive in the further development.

5.3 Cross-Evaluation of Collected Reformulations

In Alexa Developer Console, we can create an annotation set by uploading a CSV file containing a set of natural language utterances. It is possible for each utterance to specify its expected triggered intent as mandatory, expected slot types and expected slot values as optional. We utilize the NLU evaluation tool to test the natural language understanding (NLU) model for the Alexa skill with these custom annotation sets. In other words, the NLU evaluation tool aims to test whether the natural language utterances are mapped to the expected intents with the expected slot types and slot values. We select three representative intents with more qualified and fewer misunderstanding or ambiguous sample utterances to present the idea of cross-evaluation. For each of these intents, we remove the sample utterances from one of the scenarios and test whether Alexa can assign the corresponding intents to the natural language utterances from all scenarios exactly. A screenshot of an evaluation batch can be seen in Figure 5.1. We omit the specification of expected slot types and slot values but focus on the intents. The red part means the actual intent fails to be consistent with the expected intent. In this way, we can evaluate the sample utterances' generalizability and figure out the unique patterns in specific scenarios. The descriptions of all intents mentioned in the following are shown in Table 5.1. The collection types (argument, book, news or trip) are replaced by a slot type {collection} including all possible values.

Task 2 Create(l_1) with CreateQueryIntent

(1) With the sample utterances from the argument scenario removed, some unique annotations in the argument scenario fail to have the expected intent CreateQueryIntent: *what do you think about* l_1, *what are the pros and cons of* l_1, *what are the arguments for and against* l_1. As Alexa could not find the mapping sample utterances of CreateQueryIntent, they are assigned to other intents like CreateNegLiteralIntent and UpdateQueryIntent. (2) By removing the sample utterances from the book scenario, some failed annotations are: *please recommend me some books on* l_1, *please give me list of those books which describe about* l_1, *can you list books about* l_1, *could you look up some books on* l_1. Since these expressions like recommend, which describe about, list, look up rarely



< Go Back

NLU Evaluation / 37140832-3f... Source: TASK2 (aadef5ed-c9cb-4043-8ce5-771a3bf932f7)

0% Passed ()
Failed (275)
Total Test Cases: 275

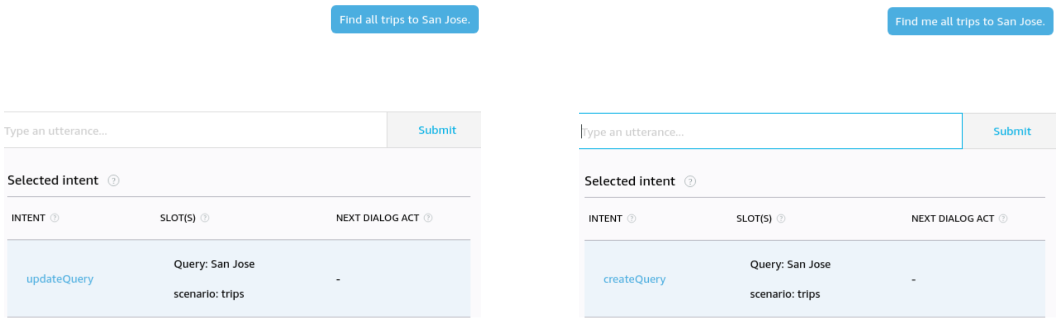
ASK_SAA_EVALUATION_RESULTS_ENDED_LABEL
17 minutes ago in **DEVELOPMENT**
Sep 12, 2020 7:48 PM

Export JSON

#	UTTERANCE	EXPECTED [E] VS ACTUAL [A] INTENT	EXPECTED [E] VS ACTUAL [A] SLOT NAMES	EXPECTED [E] VS ACTUAL [A] SLOT VALUES	PASS/FAIL
256	Show me all trips to San Jose	E: createQuery, A: updateQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
257	Show me trips to San Jose	E: createQuery, A: updateQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
258	list trips to San jose	✓ createQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
259	Can you open up all trips to San Jose	E: createQuery, A: updateQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
260	I want you to get all trips to San Jose	E: createQuery, A: updateQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
261	Can you find me all trips to San Jose please.	✓ createQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
262	Trips to San Jose.	E: createQuery, A: confirmLiteral	E: [ ], A: QueryTwo E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL
263	Show me trips to San Jose.	E: createQuery, A: updateQuery	E: [ ], A: Query E: [ ], A: scenario	E: [ ], A: San Jose E: [ ], A: trips	FAIL

Figure 5.1: Cross-Evaluation of CreateQueryIntent (Task 2) without sample utterances from the trip scenario in Alexa Developer Console.

exist in other scenarios. Interestingly, the annotation *find all books about* l_1 is failed. However, it exists the pattern find {collection} about l_1 in the sample utterances from the other three scenarios. Hence, it reveals the drawback of directly using the natural language utterances as sample utterances rather than the summarized patterns since Alexa is sensitive to the minor difference in the sample utterances. (3) By removing the sample utterances from the news scenario, *I want all news items about* l_1 is selected as UpdatePartFieldIntent instead of CreateQueryIntent, although a pattern I want all {collection} about l_1 exists. If items is removed from the annotation as *I want all news about* l_1, it is assigned with the expected intent. That is to say, it is necessary to add news items as the synonyms of news in the values of slot type. *please find me all news articles concerning* l_1 fails since concerning exists only in a pattern with a different structure once. *please show some latest news articles about COVID-19* is failed due to latest, which is unique to the news scenario. (4) Without the sample utterances of the trip scenario, 39 of 59 natural language utterances with 66% in the trip scenario are failed. This is a relatively high failure rate. The failed annotations mainly are: *show all trips to* l_1, *what trips are available to* l_1, *I would like to take a trip to* l_1, *hello I like to plan a trip for* l_1, *would you help me to do that*. These expressions such as take a trip, what trips are available are more likely to be applied only in the trip scenario. Besides, the preposition to is unique to the trip scenario as l_1 is a destination. Thus, even



Find all trips to San Jose.

Type an utterance... Submit

Selected intent

INTENT	SLOT(S)	NEXT DIALOG ACT
updateQuery	Query: San Jose scenario: trips	-

(a) *find all trips to San Jose* with UpdateQueryIntent

Find me all trips to San Jose.

Type an utterance... Submit

Selected intent

INTENT	SLOT(S)	NEXT DIALOG ACT
createQuery	Query: San Jose scenario: trips	-

(b) *find me all trips to San Jose* with CreateQueryIntent

Figure 5.2: Different selected intents for two similar annotations in Task 2. The one (a) without *me* is assigned with UpdateQueryIntent, while the other one (b) with *me* is assigned with the expected intent CreateQueryIntent. The evaluation is done by using the Utterance Profiler tool provided in Alexa Developer Console.

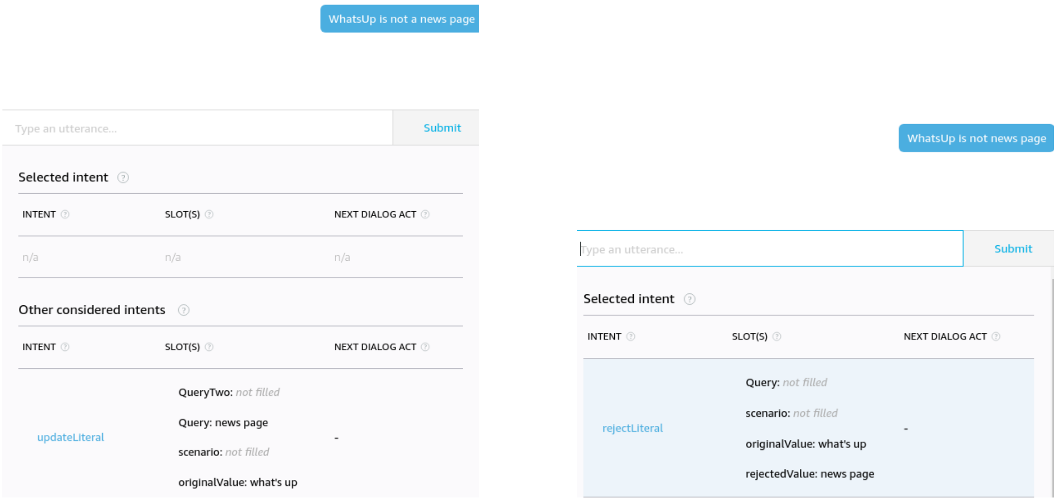
if the pattern find all {collection} about l_1 exists in the sample utterances of this intent, the annotation *find all trips to* l_1 is still failed. But this is not absolute, as the annotation *find me all trips to* l_1 passes the test. It is most likely because find me all ... is more common in the sample utterances even if the prepositions do not include to. A comparison of the selected intents for these two similar annotations is presented in Figure 5.2 by using the Utterance Profiler tool provided in Amazon Alexa Developer Console.

Task 7 Update($l_5, ?$) with RejectFilterIntent

In this intent, there are one mandatory slot type {rejectedValue} and three alternative slot types {collection}, {originalValue} and {field}. {rejectedValue} is represented as {?} and {originalValue} is represented as l_5 in the following. The failed annotations are unique to their scenarios so that the system is not able to find corresponding sample utterances from the other three scenarios when these annotations are removed from the sample utterances. (1) By taking away the sample utterances from the argument scenario, the failed annotations can be *you misunderstood me as I do not need the {?}* and *this is the wrong* l_5, *which is a {?}*. misunderstood only exists in the pattern hey you misunderstood l_5 does not mean {?} from the other three scenarios and the wrong l_5 only occurs in the argument scenario. Interestingly, the annotation l_5 *is not a {?}* is failed to have RejectFilterIntent, even if the pattern l_5 is not {?} appears in the sample utterances from the other three scenarios ten times. The annotation l_5 *is not {?}* can be assigned with the expected intent. The test for these two similar annotations can be seen in Figure 5.3. In Figure 5.3(a), although the system does not decide a selected intent, the preference of other considered intents is UpdateLiteralIntent rather than RejectFilterIntent. It shows again that only one word (e.g., "a") difference between human annotation and sample utterance might result in the different selected intents. (2) By extracting the sample utterances from the book scenario, some failed annotations are like *when I say* l_5 *I am not talking about {?}*, *I did not mean* l_5 *to translate to {?}*, and *I am not looking for books about {?}*, though it exists similar but not equivalent patterns from other scenarios as I am not talking about the {?} l_5, I did not mean l_5 as {?}, and I am not looking for a {?}. That is to say, the number of slot types between annotation and sample utterances should be consistent. Otherwise, the intent can not be mapped correctly. It is the same for the prepositions. (3) With sample utterances from the trip scenario, *that is incorrect, it is not {?}* and *that is not {?} I was referring to* are failed, while the patterns {?} is incorrect and I was not referring to {?} exist in other scenarios, which are similar to the failed annotations but just switch positions of certain words. (4) The news scenario has more sample utterances. For such tasks that have multiple requirements and complex information needs, more unique annotations exist. For instance, l_5 *is not {?}. do you get it?* and l_5 *does not stand for {?}, please correct that for me.* Despite the existing patterns l_5 is not {?} and l_5 does not stand for {?} are quite common in other scenarios but they do not contain the additional parts like *do you get it* and *please correct that for me.* The annotation *{?} is not correct* is different from the existing pattern *{?} is incorrect.* Only one participant from the news scenario uses the annotation *I am not asking for {?}.*

Task 10 Read(Q_{10}) with ReadQueryIntent

There are 189 natural language utterances in the annotation set for Task 10 with 41 from the argument scenario, 45 from the book scenario, 37 from the trip scenario, and 66 from the news scenario. The slot type {collection} can be embedded in the sample utterances of ReadQueryIntent as optional. In general, the sample utterances of this intent have better generalizability. That is to say, with the sample utterances from any three scenarios, most annotations are mapped into the expected intent ReadQueryIntent. There are only a few exceptions. (1) With the sample utterances from the argument scenario removed, *please tell me about my previous query* and *please confirm my argument criteria.* previous query is commonly used in other scenarios, while it is embedded in the structures like tell me what I ... and tell me [my|the] ... instead of tell me about .... Besides, confirm only occurs in the pattern can you please confirm the information I provided earlier in the other three scenarios and {collec-



WhatsUp is not a news page

Type an utterance... | Submit

Selected intent

INTENT	SLOT(S)	NEXT DIALOG ACT
n/a	n/a	n/a

Other considered intents

INTENT	SLOT(S)	NEXT DIALOG ACT
updateLiteral	QueryTwo: *not filled* Query: news page scenario: *not filled* originalValue: what's up	-

(a) *WhatsUp is not a news page* without selected intent

WhatsUp is not news page

Type an utterance... | Submit

Selected intent

INTENT	SLOT(S)	NEXT DIALOG ACT
rejectLiteral	Query: *not filled* scenario: *not filled* originalValue: what's up rejectedValue: news page	-

(b) *WhatsUp is not news page* with RejectFilterIntent

Figure 5.3: Different selected intents for two similar annotations in Task 7. The one (a) with *a* fails to be assigned with any intent, while the other one (b) without *a* is assigned with the expected intent RejectLiteralIntent. The evaluation is done by using the Utterance Profiler tool provided in Alexa Developer Console.

tion} criteria only occurs in the pattern what was the {collection} criteria for my previous search once. Hence, Alexa fails to find the corresponding patterns for these two annotations. (2) By removing the sample utterances from the book scenario, some annotations that include the collection type books are failed: *please tell me which ideas I gave you about the kind of books this list shows me* and *what is the search query including for the book list?*. (3) By removing the sample utterances from the trip scenario, similar to (2), *what specifications are the trip I am looking at* and *can you let me know the previous rule I gave you as far as bringing up trip information* are failed to be selected as ReadQueryIntent. For (2) and (3), the failed annotations are unique to their belonging scenarios. (4) If the sample utterances from the news scenario are removed, there are more failed annotations: *repeat previous query*, *display last query*, *repeat last query*, *please list all search terms and parameters which are currently being applied*, and *repeat directions to populate list news*. In fact, this problem can be solved if the organized patterns are allowed to implement, as the annotations *please tell me about my previous query* and *repeat last search please* from the other three scenarios can be summarized as the pattern [please|] [tell me about|repeat] [my|] [previous|last] [search|query] [please|]. The annotations *repeat previous query* and *repeat last query* are included in the pattern such that they can be identified as the expected intent ReadQueryIntent.

Chapter 6

Results

In the utterances labeled as "good", multiple utterances are summarized into a template because of the similar syntax. The templates that have multiple interpretations (also refer to ambiguity) are discarded in the patterns of each task. We present the prevailing patterns of each task in different scenarios (argument, book, news, and trip) in Section 6.1 and the patterns commonly used in different countries (Australia, Canada, India, the United Kingdom, and the United States) in Section 6.2. The summary of the analysis of various reformulation patterns are detailed in Chapter 7.

In the summarized templates, [A|B] represents that it could be either A or B in this position. [C|] represents that it could be either C or nothing here. {collection} denotes a set of all collection types [arguments|books|news|trips] and {field} denotes a set of specified fields of different collections [sources|genres|headlines|hotels]. The complete templates with the frequencies mentioned in this Chapter can be seen in Table A.1. Task 1 is a 'ready' task, which intends to leave time for participants to read the instruction of the study and get familiar with the highlighting strategy of keywords with underline. Task 2 to Task 13 correspond to the operations in Table 3.2.

6.1 Patterns in Scenarios

This section presents the common patterns of each task with different intents in four topic-oriented search scenarios (argument, book, news and trips) and some unique patterns in certain scenarios. The collected patterns of each task are presented in a single subsection, except Task 5 and Task 11. In these two tasks, the intents of a significant number of participants' answers are different from the expected intents of the tasks. Thus, they are split into two subsections, where Task 5a (Page 51) and Task 11a (Page 55) include the patterns with the expected intents and Task 5b (Page 52) and Task 11b (Page 55) contain the patterns with the newly introduced intents. Meanwhile, we perform an evaluation to figure out the overlapping patterns of different intents, shown in an independent Subsection Overlapping Patterns.

Task 2 Create(l_1): start a search for main-topic

Participants are used to using a command [please|] [show|find|get|give|] [me|] [a list of|all|] {collection} [on|about|related to|to] l_1 to start a session for searching for a collection about a certain main-topic, followed by [can|could] you [please|] [find|get|show|give|list] [me|] [a list of|all|] {collection} [on|about|to] l_1. Utterances that omit verbs are also sorted into the first pattern. can I [have|see] [a list of|all|] {collection} [on|about|to] l_1 and I [am looking for|want|would like] {collection} [on|about|to] l_1 are the next two popular patterns. In the trip scenario, the connection words between the required collection type trips and l_1 mainly are [to|for] rather than [on|about], since the main difference from other scenarios is that l_1 is a destination San Jose. This feature is also revealed in other tasks. There are some special patterns as I like to plan a trip for l_1 and what trips are available to l_1. In the argument scenario, the conjunction between arguments and l_1 can be for and against besides [on|about]. The alternative noun for arguments is pros and cons, existing in utterances such as *I am looking for the pros and cons of banning plastic bags* and *What are the pros and cons of banning plastic bags*. An implicit way to ask for arguments is: what do you think about l_1.

Task 3 Create(l_2): add subtopic to filter results

Participants are expected to add a condition l_2 to filter the results based on the main-topic l_1. Although a large number of utterances meet the task requirements, however, meanwhile they are ambiguous themselves. That is to say, there are multiple interpretations for these utterances, especially in the context of the whole session. For instance, [just|only|] show [me|] [just|only|] {collection} about l_2, whether the utterances do not include [just|only], or include [just|only] either before verbs or after verbs but before collection types, they can also be interpreted as Update(l_1, l_2), which indicates the results should only include collections about l_2 that have nothing to do with l_1. Interestingly, [only|just] in different positions might result in different explanations. For example, can you show me trips that are only by ship in the trip scenario where l_2 is by ship, it can also represent the operation Update($l_1, l_1 \vee l_2$) and indicate the list includes trips by ships, regardless of the destination is San Jose (l_1), in addition to the trips to San Jose. These ambiguous templates are excluded in the patterns of

Task 3.

There are mainly three categories in the unambiguous patterns of Task 3. (1) use pronouns: [can you|] [only|just|] [show|give|include] [me|] [only|just|] [the|] [ones|those] [that are|] [about|on] l_2, which [of|] [these|ones] [include|relate to|are] l_2. The usage of pronouns generally causes ambiguity in a sentence without contextual information, which is known as coreference phenomenon in conversation. Nevertheless, this problem can be solved in our system as it is able to figure out what the pronouns refer to by referencing the context of the whole dialogue. Here, the pronouns [ones|these] specify the collection in the result list returned last turn. With these pronouns, the system will know the next operation should be performed based upon the previous results. The connection word between pronouns and l_2 are is unique to the trip scenario. In the argument and news scenarios, there is a pattern that directly uses results as [could you|] [only|] show [me|] results that [include|also mention|are about] l_2. Other pronouns like them and there are also used: [please|] [give|show] me [arguments|books] [that focus on|about] l_2 [from|in] them [only|] and I want to have a travel to there l_2. (2) use verbs like filter, trim down, narrow down, reduce, shorten: [can you|] [filter|shorten|trim down|narrow down|reduce] [these|that|list|{collection}] [to|] [those|ones|] [only|just|] about l_2, filter [these|this list] [with|for] l_2 [only|]. Participants will use these verbs only when they receive something as results previously and want to process these results with certain operations. So it is clearly that the new results are first related to l_1 and then are also about l_2. (3) elimination: [please|] [remove|filter out] [all|] {collection} [that do|that are] not [relate to|about|] l_2. This is an indirect way to satisfy the task requirements and only a few participants use this pattern in all scenarios (with 1 in argument, 3 in book, 2 in news, 1 in trip).

Task 4 Update($l_2, l_2 \vee l_3$): add alternative subtopic

Participants are expected to add another condition l_3 as an alternative to l_2. The results should contain collection either about l_1 and l_2 or about l_1 and l_3. Similar to Task 3, there are many ambiguous templates. The most popular templates are: [please|] [also|] [add|add back|include|expand] [to the list|] {collection} [that|] [on|about|are|mention|talk about] l_3 [to the list|] and [can|could] you [please|] [also|] [add|include] {collection} [about|including|containing|] l_3 [to the list|] [as well|]. These templates can also be interpreted as the operation Update($l_1 \wedge l_2$, ($l_1 \wedge l_2$) $\wedge$ l_3), which means the result list should include the collection only about l_3 besides the collection about l_1 and l_2. That is to say, the system confuses if participants want to have the collection about l_3 that is also related to previously mentioned l_1, or l_3 is another independent main-topic keyword. In addition, for templates [show|fetch] [me|] {collection} [about|related to|containing] l_3 and [can|could] you [tell|get|show] [me|] {collection} [that are|] [on|related to|about] l_3, which are the same as the patterns in Task 2, so they can also represent the operation Update($l_1 \wedge l_2$, l_3), where replaces collection about l_1 and l_2 with collection about l_3. The patterns of Task 4 will not cover these ambiguous templates.

The unambiguous patterns of Task 4 can be classified into three categories: (1) use pronouns: can you [also|] [add|show me|include] [those|ones] [including|about|that discuss] l_3 [too|as well|], [now|] [show me|include] [the|] [ones|those] about l_3 [too|as well]. With these pronouns, the system goes back to the context in earlier interactions to find what the pronouns are referring to and add conditions l_3 based on what pronouns refer to. In the book scenario, there is a pattern that without pronouns but uses other expression to indicate the already mentioned concept: could you also suggest me something on the same topic but for l_3. (2) mention both l_2 and l_3: can you [find|search|open|get] [me|] [all|] {collection} [about|on] l_2 or l_3, [now|] [please|] [find|show|list] [me|] [all|] {collection} [that are|] [about|] [either|] l_2 or l_3, filter with either l_2 or l_3, I [am looking for|want to see] {collection} about either l_2 or l_3. Since the system already knows l_2 is a subtopic of main-topic l_1 last turn, mentioning both l_2 and l_3 with conjunction 'or' is to specify the correlation between l_2 and l_3. There are some unique patterns that are proposed by only one participant in a certain scenario. In the argument scenario: are there any arguments about l_3 as an alternative to l_2 and I would like arguments related to l_2 as well as those arguments related to l_3. In the book scenario: could you change it to l_3 or l_2, the pronoun it could refer to the filter condition l_2. In the news scenario: please gather all news about l_3 that would be an alternative to l_2. In the trip scenario: if there are no choice to l_2 transport, will alternatively show me the trip l_3 and show me trips l_2, if ship trips not available then I would like to select trips l_3. Both patterns are likely only applicable to the trip scenario. (3) l_3 is an alternative filter: [can you|] [please|] add l_3 [to filter|] as [an|] alternative? and [or|another option is] l_3.

For the query Q_4 as $((l_1 \wedge l_2) \vee (l_1 \wedge l_3))$ or $((l_1 \wedge l_2) \wedge (l_1 \wedge l_3))$, the system will give the same results in some cases. For instance, the system can find collection about l_1 and l_2 and collection about l_1 and l_3 as well, even if seekers do not specify both are mandatory, but it is always good to return both as results to satisfy the seekers' information needs beyond their expectations. On the other hand, if seekers require both, but the system can only find one of them, it is still necessary to return the found part as a result rather than nothing. Thus, in this regard, we do not strictly distinguish between both queries.

Task 5a Delete($l_1 \wedge l_2$): remove part of filters

In Task 5, the filter conditions l_2 and l_3 are expected to remove. The results only include collection about l_1, the same as the results after Task 2. Unlike Task 3 and Task 4, many utterances are not ambiguous themselves but participants misunderstand the task and mix up the intents of updating part and deleting part. Deleting part refers to release the filter conditions, however, updating part refers to replace the existing literals with the negation of these literals (Update($l_2 \vee l_3, \neg l_2 \vee \neg l_3$) or Update($l_2 \vee l_3, \neg l_2 \wedge \neg l_3$)). The patterns with the intent of updating part are classified into a separate Task 5b. Hence, there are only a few patterns for Task 5a. The patterns can be divided into the following types: (1) add back collection not about l_2 and l_3: [can you|expand the list to|] [add|include|also tell me] [back|] [all|any|] [{collection}|ones] [that are|] not [about|related to|] l_2 or l_3 [to the list|]. This pattern appears most frequently in the argument and book scenarios. Interestingly, the existence of the word also and its position can result in different meanings of a utterance. For example: can you tell me {collection} that are not about l_2 or l_3 and can you also tell me {collection} that are not about l_2 or l_3. The second pattern is more likely to ask for collection not about l_2 or l_3 as additional supplements on the original results. (2) show collection [other than|not just] l_2 and l_3: can you [include|find|show|provide|get] [me|] [all|] {collection} [in addition to|other than|not just] [the ones|] [regarding|] l_2 [and|or] l_3 and [find|show|include] [me|] [all|] {collection} [other than|besides|not just] l_2 [and|or] l_3. These patterns are most popular in the news and trip scenarios. Compare the argument scenario and the trip scenario, [other than|besides|including] l_2 and l_3, [what are other factors at play|what are other advantages and disadvantages|are there other reasons] from the argument scenario, factors, advantages and disadvantages, reasons are exclusive in the argument scenario, while options is exclusive in the trip scenario, used in the pattern besides l_2 and l_3, what other options do you have in order to get there. (3) remove filter l_2 and l_3: It exists in the book, news and trip scenarios. remove l_2 and l_3 from filters in the book scenario, please remove filters [on|] l_2 and l_3 and can you remove l_2 and l_3 filters in the news scenario, remove filters l_2 and l_3 [and show me all trips|] in the trip scenario. A similar pattern in the argument scenario: forget about only including l_2 and l_3.

Task 5b Update($l_2 \vee l_3, \neg l_2 \vee \neg l_3$) or Update($l_2 \vee l_3, \neg l_2 \wedge \neg l_3$): update part of filters to negation of filters

Most utterances are adopted in the argument, book and news scenarios with 35, 39 and 58 respectively. There are 28 utterances from the trip scenario. The pattern [now|] [please] [find|show] {collection} [that are|] not [about|] l_2 or l_3 is widely used in all scenarios, which might be affected by the description of the task: *but now again with those {collection} that are not about l_2 or l_3* (Table 4.1). It also found that participants mix up two kinds of expressions in their utterances: one is like not [related to|include|use|have|about|] l_2 or l_3, [exclude|without|except|remove] l_2 or l_3, representing the operation Update($l_2 \vee l_3, \neg l_2 \vee \neg l_3$); the other one is like [exclude|remove|avoid|except|filter out|without] l_2 and l_3, neither l_2 nor l_3, not l_2 and l_3, representing another operation Update($l_2 \vee l_3, \neg l_2 \wedge \neg l_3$).

Task 6 Create($f_1 : (l_4 \wedge l_5 \wedge l_6)$): specify certain field with part of filters

In the argument and book scenarios, the most commonly used pattern is: [only|] [show|give] [me|] {collection} [that are|] [from|with] [one of|] [the|] [following|] [sources|genres] l_4 [or|,] l_5 or l_6, followed by [can|could] you [please|] [just|only|] [show|find] [me|] {collection} that are from [sources|genres] l_4 [or|,] l_5 or l_6. However, in the news and trip scenarios, participants are used to mentioning keyword conditions before field name, for example, the most popular pattern is: [please|] [just|only|] [show|give] [me|] [all|] {collection} [that|which|] [with|include|contain|have] [either|] l_4 [or|,] l_5 or l_6 [as the hotels|in the headlines]. Participants perform well in this task and have various ways to describe the collection with the specific field, especially in the trip scenario such as trips that are book at ... hotels, trips where I stay at the following hotels: ..., include accommodation in ... hotels. The real utterances are more complex than the patterns shown here. The complete patterns in each scenario can be seen in Table A.1. An indirect way to satisfy the task requirements is to remove the collection that are not l_4, l_5, or l_6 in the field. Such a way is more popular in the book scenario as: [please|] remove books that [are not|do not have] [in the|] l_4 [or|,] l_5 or l_6 [in the|] genre [name|]. In each argument, news and trip scenarios, one participant uses a similar pattern like [please|] remove [from the list|] {collection} where [source web page|headline|hotel] [is not|does not contain] [one of|] l_4 [or|,] l_5 or l_6.

Task 7 Update($l_5, ?$): indicate error and reject filter

This task aims to ask participants to give some negative feedback when they figure out the chatbot's recognition error. Participants are required to point out the error and reject the unwanted value without telling the expected value in their utterances. In all scenarios, l_5 [that I meant|] is not [equal to|] $\{?\}$, l_5 does not [mean|refer to|stand for] $\{?\}$ and [No,|] I [did not mean|was not referring to|was not looking for] [l_5 as|] $\{?\}$ have high frequencies, where $\{?\}$ refers to the rejected value. There are only a few participants in each scenario using $\{?\}$ is incorrect. Except for the argument scenario, I [did not mean|was not referring to] $\{?\}$ [when I said|by|for] l_5 exists in the other three scenarios. Nevertheless, other popular patterns only indicate a certain value is unwanted, but without specifying that it is an error made by the system. For instance, [please|] [remove|exclude|take out|omit] [all|any|] {collection} [referring to|about|with|that reference] $\{?\}$, do not [show|bring|give|include] [me|] [{collection}|] [that include|] $\{?\}$. The key point is whether seekers do not want the filter for their own reasons, or the filter is wrongly interpreted by the system so that they do not want the unexpected filter. For the latter, the system must apologize and confirm with seekers using *What do you mean by* l_5. Thus, the utterances that do not reveal the error can be problematic. This task is worth comparing with Task 13 that represents the operation Create($\neg\ l_1$), as it exists the expression like not ... in the patterns of both tasks. The comparison results are elaborated in Subsection Overlapping Patterns.

Task 8 Update($?, l_7$): ask for confirmation

After the system recognizes its error, the system has to apologize and confirm the participants' real expected value. Then the participants inform the chatbot about the correct value. The system asks a question *What do you mean by* l_5, most participants intuitively use statements like I ... to express their needs such as I [actually|] [mean|want|would like] [to|] [say|include|] l_7, I [was|am] [thinking of|looking for|talking about|referring to] l_7. Participants are also allowed to mention the original value l_5 at the same time as l_5 [actually|] [means|refers to|stands for|denotes] l_7, which is especially common in the book and news scenarios. In each book and trip scenarios, there is a pattern including the rejected value as [could you|] change [filter from|] $\{?\}$ to l_7. Interestingly, it can be found that the prevalent patterns in Task 8 correspond to the prevalent patterns in Task 7. For instance, the structure [l_5|I] [does|do] not mean ... in Task 7 to reject the unexpected value and [l_5|I] mean ... in Task 8 to reveal the correct value.

Task 9 Update(l_1, l_8): replace old filter with new filter

The main-topic l_1 will be replaced with another new topic l_8 but keep conditions for certain field unchanged. In all scenarios, the most frequent pattern is [now|also|only|] [show|list|give] [me|] [all|] {collection} [that are|] [about|to] l_8 [not|rather than|instead of] l_1 [please|]. The structures of other patterns are: change {collection} from l_1 to l_8, [instead of|exclude|remove] {collection} l_1, [I need|show] {collection} l_8, replace l_1 with l_8. In this task, participants are free to indicate the consistent conditions for certain field. There are 19 of 53 "good" utterances that use alternative representations to describe the same conditions in the argument scenario, 15 of 57 "good" utterances in the book scenario, 23 of 93 utterances in the news scenario, and 17 of 56 "good" utterances in the trip scenario. [from|with|using|keep] [these|the] [same|previous|] [sources|headlines|hotels] [filters|] [that I have previously mentioned|as before|] is widely used in the argument, news and trip scenarios. In the book scenario, the alternative descriptions of unchanged conditions are: [in|on|from|by] [those|the same|] genres [I mentioned|], I want to keep the previous genres. In each argument and trip scenarios, one participant uses the same criteria instead of referring to the field. Interestingly, in the trip scenario, as l_1 and l_8 are city names, some participants use [I want to|] [change|replace] [the|] [city|trip destination|destination|trip] [from|] l_1 [to|with] l_8.

Task 10 Read(Q_{10}): recall search history

Participants ask chatbot to recapitulate their search history such as query, search parameters. In formal studies, participants perform better than in the pilot studies. In 60 answers of each argument, book and trip scenarios, there are 41, 45 and 37 "good" utterances respectively. There are 66 "good" utterances of 100 answers in the news scenario. The popular patterns in these scenarios have the similar structures: [please|] [show|tell] [me|] [my|] [previous|earlier|last|] [query|requests|search|criteria|message|search history|command history|search parameters|search details|filters], followed by [sorry,|] [I|] [lost track of where we were|forgot what I asked you previously|got confused|], [can|could] you [please|] [show|tell|repeat|remind] [me|] what I [previously|] [have asked for|asked you to do|have searched for|told you] [before|previously|earlier|]? Besides what I ..., there are other expressions like what [this list is based on|filters have you currently applied|you are searching for|my previous inquiries were]. The verbs adopted in other patterns: repeat, confirm, find, remind, recap.

Task 11a Update($f_1 : (l_4 \vee l_7 \vee l_6), l_9$): replace part of filters with new filter

The filter conditions for certain field are expected to remove and a new filter l_9 for the general search is added. However, similar to Task 5, many participants misunderstand this task. Some of them specify l_9 as one of the conditions for certain field and the others confuse with the operation Update($f_1 : (l_4 \vee l_7 \vee l_6), \neg(f_1 : (l_4 \vee l_7 \vee l_6)) \wedge l_9$). The patterns with this operation are sorted into an additional Task 11b. In order to remove restrictions of a field, in the argument and trip scenarios, es exists I do not care about the [source|hotel], while in the book and trip scenarios, es exists the [genre name|hotel] does not matter. The common expressions in all scenarios are: [remove|ignore|cancel|clear|delete] [the|all|any|] {field} [filters|parameters|requirements|constraints|instructions| conditions|keywords] and [without a specific|have nothing to do with|pay no attention to|from any|with any] {field}. In fact, more participants do not realize the notion of filters, requirements or constraints, so that they use [instead of|not|rather than] [a|the|] {field}.

Task 11b Update($f_1 : (l_4 \vee l_7 \vee l_6), \neg(f_1 : (l_4 \vee l_7 \vee l_6)) \wedge l_9$): update part of filters to negation of filters and add a new filter

14 utterances of 60 answers are adopted in each argument and book scenarios and only 7 utterances of 60 answers are from the trip scenario. In the 100 answers of the new scenario, there are 14 utterances representing this intent. Hence, compared to the argument and book scenarios, fewer participants misunderstand the task in the news and trip scenarios. In the argument scenario, they describe the conditions for source as: [not|without] from [any|] [previous|] source [previously mentioned|from before], [not a source from|different than] l_4, l_5 or l_6. In the book scenario, the descriptions of conditions for genre names are like: [instead of|without] [in the|] [one of|] [those|] genre names [that I mentioned before|l_4, l_5 or l_6], remove {collection} with genre names l_4, l_5 or l_6.

Task 12 Update(Q_{11}, l_{10}): replace entire query with new filter

The whole query Q_{11} after Task 11 is changed to a new query containing only l_{10}. It is interesting to see how many participants will clearly point out a completely new search. In the results of these scenarios, there are 21 utterances of 56 "good" utterances with 37.5% indicating the new search in the argument scenario, 32 of 58 (55.2%) in the book scenario, 42 of 102 (42%) in the news scenario and 24 of 53 (45.3%) in the trip scenario. There are mainly two ways to emphasize the new search: (1) [forget|ignore|remove|clear] [all|] [previous|] [everything|search requirements|criteria|parameters|commands|instructions|search filters|search criteria|list|queries|requests|searches], (2) [let us|] [start|] a new [search|list|request|query] [for|], I have a new request and show me a [completely|brand|] new list. However, if the participants do not highlight the search should be a completely new one, the utterances are almost the same as the utterances in Task 2 (Create Query). The overlapping patterns that lead to the misidentification of both intents are presented in the Subsection Overlapping Patterns.

Task 13 Create($\neg$ l_1): add negation of filter to query

Since we use l_9 in the book and trip scenarios by accident, [l$_1$|l$_9$] is used in the summarized patterns. The most frequently used patterns are similar in the argument, book and news scenarios: [remove|filter out|exclude|without|discard|eliminate] {collection} [that are|] [about|related to|on] [l$_1$|l$_9$] [from the list|]. The second common pattern in the book scenario is [get|show|include|] [me|] [only|] {collection} [that|which] are not [about|related to] l$_9$, which is different from the argument scenario. The second popular pattern in the argument scenario has the similar structure [remove|exclude] ... l$_1$, but starting with [can|could] you. However, in the trip scenario, participants are more used to saying [please|] [exclude|without|remove|eliminate|forget|discard] l$_9$ [trips|from trips|from filters|filters|] [from the list|], followed by [please|] [show|search] [me|] [all|only|] {collection} [that|which|] [do not involve|are not about|without|but exclude|not including] l$_9$. As we mentioned earlier in Task 7, the patterns in Task 7 and Task 13 may overlap and the evaluation results are introduced in the next Subsection Overlapping Patterns.

Overlapping Patterns

After building the query interaction model (Chapter 5.1), we investigate if Amazon Alexa is able to exactly distinguish the patterns of different intents by evaluating Alexa's accuracy in assigning various utterances to the corresponding intents. The results show that Alexa will mix up the intents of utterances that represent identical patterns between Task 2 and Task 12 and between Task 7 and Task 13. For Task 2 and Task 12, the misidentified utterances can be: *Can you get me all new articles about COVID-19* and *Show trips to Santiago.* But Alexa can assign the utterances of Task 12 that have tiny differences from Task 2 to the corresponding intents correctly. That is to say, Alexa is sensible to these minor differences between the patterns. For example, [bring|show] [me|] {collection} about l$_{10}$ [instead|now], [show|find|give|tell] [me|] [different|only|just|instead] {collection} [that include|about|for] l$_{10}$, [just|only|now|instead] [tell|show] [me|] {collection} [relating to|about] l$_{10}$, I now only want {collection} [about|to] l$_{10}$. This can also reflect that seekers generally will not use instead, just, only, now, different when they first create a query in a search session. For Task 7 and Task 13, an annotation set, consisting of a total of 417 utterances from both tasks, is used for the evaluation. The results revealed that it indeed exists overlapping patterns between both intents, and a total of 29 utterances that should be recognized as createNegLiteralIntent (as expected in Task 13) were matched to the wrong intent rejectLiteralIntent (specified for Task 7). The misidentified utterances mainly occurred in the book, news and trip scenarios with 10, 7 and 8 utterances respectively, while only 4 utterances are misidentified in the argument scenario. The overlapping patterns are mainly: I do not want [l$_1$|l$_9$], do not [show|include] [me|] [any|] [l$_1$|l$_9$], [can you|] show me [any|] {collection} that are not about [l$_1$|l$_9$], [exclude|remove] [l$_1$|l$_9$].

6.2 Patterns in Countries

In this section, we advance the common patterns of each task with different intents under the news scenario in five countries (Australia, Canada, India, the United Kingdom and the United States) and some unique patterns in certain countries. With the exception of Task 5 and Task 11, each other task's results are shown in a subsection independently. For Task 5 and Task 11, the intents of a large number of participants' answers vary from the expected intents of the tasks. As a result, they are divided into two subsections, where Task 5a (Page 58) and Task 11a (Page 61) cover the patterns with the expected intents and Task 5b (Page 59) and Task 11b (Page 62) incorporate the patterns with newly introduced intents.

Task 2 Create(l_1): start a search for main-topic

In all countries, there is a most popular pattern: [please|] [show|find|get|tell] [me|] [all|some|a list of|] [latest|] news [on|about|related to] l$_1$. Some participants omit the verbs in the utterances with the pattern [all|latest|a list of|] news [about|on] l$_1$ except ones from Canada. This pattern is the second common pattern in Australia and India. In Canada and the United Kingdom, participants are more used to saying [can|could] you [please|] [get|find|show] me news [on|about] l$_1$. A few participants prefer using [list|give|find] [me|] [all|any|latest|] l$_1$ news rather than news [about|on] l$_1$, especially in the United States. But no Indians use this expression l$_1$ news. The British participants use a variety of prepositions to connect news and l$_1$ such as [on|concerning|about|related to|regarding|which cover], while participants from other countries mainly use [on|about|related to].

Task 3 Create(l_2): add subtopic to filter results

As mentioned in Section 6.1, many templates that are ambiguous themselves are discarded in the patterns of Task 3. In the news scenario, most unambiguous patterns are from Canada, whereas fewer from the United States. In other countries except Canada, most adopted patterns include pronouns to indicate the previous results such as [can you|] [just|only|] show [me|] [only|just|] [the|] [ones|those] [about|related to|that discuss] l_2. Two patterns in Canada use pronouns: please filter articles to ones that contain l_2 and I want it to contain only news about l_2, where it represents the previous obtained list. However, in a pattern show more details about its l_2 of Australia, its represents the main-topic COVID-19's. The participants from Canada prefer [filter|keep] articles [to|] [about|related to|containing|mention only] l_2. In Canada and the United Kingdom, a few participants use the implicit way to satisfy the task requirements as [filter out|remove] [all|] articles [that|which] do not [cover|mention] [subject|] l_2. Other patterns in Canada with the intent of revising list are like narrow the list down to articles about l_2 and reduce list to only show l_2 articles.

Task 4 Update($l_2, l_2 \vee l_3$): add alternative subtopic

Similar to Task 3, a large number of utterances are ambiguous such that they are removed from the patterns of Task 4. Fewer utterances are adopted as patterns in India and the United Kingdom. In Australia, Canada and the United States, most patterns contain both l_2 and l_3 as l_2 or l_3 articles in Australia, can [I|you] [just|] [see|get me] news [about|on] l_2 or l_3 in Canada, and I want to see news about either l_2 or l_3 in the United States. A few participants from India and the United States prefer using pronouns ones in the pattern [can you|] [also|] include ones [about|that discuss] l_3 [as well|]. One participant from Australia gives an uncommon pattern please gather all news that are about l_3 that would be an alternative to l_2 and another unique pattern in Canada is alter previous filter and keep articles containing l_3.

Task 5a Delete($l_1 \wedge l_2$): remove part of filters

The utterances that are misunderstood as negating part of filters instead of releasing part of filters as expected are organized into the patterns of Task 5b. In all countries, the most common patterns are those include such expression news [not just|other than|apart from] l_2 [and|or] l_3 in their utterances. These patterns are: [show|get] [me|] [all|] news [other than|not just|apart from] [the ones related to|those about|] l_2 [or|and] l_3 [ones|], which is from all other countries except Canada, can you [please|] [get me|change|list] [all|] articles [to show information|] [not just|other than] l_2 or l_3, which exists in Canada, the United Kingdom and India, and can I see more articles that show information other than l_2 or l_3 only exists in Canada. Participants from Australia and the United Kingdom specify remove filters like can you remove l_2 and l_3 filters in Australia and please remove filters on l_2 and l_3 in United Kingdom. Patterns that ask for bringing back the articles that are not about l_2 or l_3 to the previous result list only exist in the United States. For instance, include articles that do not include l_2 or l_3 too and [actually|] also [add|get] articles [that are|] [about|related to] l_2 or l_3.

Task 5b Update($l_2 \vee l_3, \neg l_2 \vee \neg l_3$) or Update($l_2 \vee l_3, \neg l_2 \wedge \neg l_3$): update part of filters to negation of filters

There are a total of 58 misunderstanding utterances from 5 countries in the news scenario such that they are summarized into the patterns of Task 5b. Most participants who misunderstand the tasks are mainly from India, followed by Canada and Australia. The most common pattern in all other countries except the United States is [now|] [please|] [show|tell|find] [me|] articles [that are not|that do not|not|except|excluding] [about|on|discuss|including|related to|have] l_2 or l_3. In the United States, the most popular pattern is I [only|] [want|would like|need] [to see|] [more|] news [but|] [that are|] not about l_2 or l_3. Above patterns can be interpreted as the operation Update($l_2 \vee l_3, \neg l_2 \vee \neg l_3$). However, it exists other patterns that represent the operation Update($l_2 \vee l_3, \neg l_2 \wedge \neg l_3$). For example, in Canada, there is a pattern [remove|exclude|take away] [articles|] [on|about] l_2 and l_3 [from original list|from the list|].

Task 6 Create($f_1 : (l_4 \wedge l_5 \wedge l_6)$): specify certain field with part of filters

In the news scenario, the specified field is the headline. It can be obviously found that most participants from all five countries prefer using expression l_4, l_5 or l_6 in the headline in their utterances. These utterances are organized into a pattern: [please|] [only|] [show|give|include|find] [me|] [just|only|] news [which|that|] [contain|have|with|include] l_4, l_5 or l_6 in the headline. Participants from Australia use various prepositions to connect news with conditions l_4, l_5, l_6 such as [which|that|] [have|mention|include|reference|contain|with|including|related to], while participants from the other four countries mainly use [that|which|]

[contain|include|with]. Only a few participants from these countries use the pattern [please|] [show|] [me|] news with headlines [related to|including|containing| mentioning] l_4, l_5 or l_6.

Task 7 Update($l_5, ?$): indicate error and reject filter

In Australia, India and the United States, more participants are used to saying l_5 [is|does] not [refer to|mean|equal to|stand for|] {?}, which clearly points out the error as well as the unexpected value. Interestingly, in Canada and the United Kingdom, the more common pattern is [please|] [exclude|remove|hide|omit] [any|] news [that contain|which mention|including|with|relating to|containing|about|] {?}, which only indicates the unwanted value without emphasizing it is a error made by the system. According to the results of the evaluation, this pattern is problematic since it will be mixed up with another Task 13 Create($\neg l_1$). Another common pattern in all five countries is I [am|did] not [asking for|referring to|mean] {?} [for|by|when I said|] [l_5|], which is similar to the pattern I did not [mean|intend for] l_5 [as|to refer to] {?}, existing in Canada, the United Kingdom and India. In Australia and the United States, there is one participant using the pattern {?} is incorrect. There is one unique pattern from the United States as the interpretation you had for l_5 as {?} is not what I actually meant. A few participants from Australia and India request for correcting the value like please correct results, l_5 does not refer to {?} in Australia and just correct l_5, does not mean {?} so correct it in India.

Task 8 Update($?, l_7$): ask for confirmation

The most popular pattern in all five countries is similar: I [actually|] [meant|mean| was referring to|want|was looking for] [l_5|] [as|to search for|for|] l_7. The second common pattern in the other four countries except India is l_5 [means|stands for|should refer to|is equal to|is interpreted as] l_7. In India, the second common pattern is [it is|] l_7, while this pattern is not so common in the other three countries except India and the United States. A unique pattern from the United States is l_5 is an abbreviation for l_7. This pattern is applicable when the system interprets the original abbreviation l_5 wrongly. For instance, in the news scenario, the abbreviation *NI* was wrongly interpreted as *North Ireland*. However, seekers meant *National Insurance*. Another unique pattern from the United States is what I had in mind initially was l_7.

Task 9 Update(l_1, l_8): replace old filter with new filter

In all five countries, the most common pattern is [please|] [show|give|find|search for|get] [me|] [all|a list of|some|] news [that are|] [about|containing|including|related to|on] l_8 [instead of|not|remove articles about|rather than|excluding] l_1. Another very popular pattern in the United States is I [want|would like|need] [to see|] news [mentioning|about|on] l_8 [and|] [not about|instead of|not on] l_1, which also exists in the patterns of Canada. There is a variety of ways to indicate the unchanged conditions for the field. [with|keep] [the same|these|those] [headlines|headline filters] [as before|] is widely used in these countries. In Australia, other options are using the same filters or in the previous headline search. In Canada, that match the criteria [I gave you|] is used, which is akin to that meet the criteria above in India and they should meet the previous criteria in the United States. In the patterns of Canada and the United Kingdom, two patterns are very similar: in addition to the previously mentioned keywords in the title, change the articles to include l_8 instead of l_1 from Canada and can you find me articles with those initials in the title but to do with l_8 [, as opposed to|but not including] l_1 from the United Kingdom.

Task 10 Read(Q_{10}): recall search history

[please|] [repeat|show|remind me of|tell] [me|] [my|the] [previous|past|last|current] [query|message|search|inputs|requests|command|filters] [criteria|details|entries|] is most frequently used in the other four countries except the United States. In the United States, the most common one is what was [my|the] [last|previous] [request|search] [query|], followed by [please|] [go back to|show me|repeat] my previous [search|request]. The second common patterns of Australia are what did I [just ask to search for|say|tell you to do before]? and what was my last [query|request]? However, in Canada, more participants prefer can you [please|] repeat [my|the] [previous|last] [information I provided|request], can you please tell me what [the latest search parameters were|I had previously requested from you] and what did I [previously ask you to do|search for]. In the United Kingdom and India, a few participants are used to saying what was my [last|previous] [search|request|message] [criteria|parameters|].

Task 11a Update($f_1 : (l_4 \vee l_7 \vee l_6), l_9$): replace part of filters with new filter

In task 11, the misunderstanding utterances that represent part of filters for a certain field are negated, and add another new filter for the general search are sorted into the Task 11b. In Australia, Canada and the United Kingdom, the most commonly used pattern has the similar structure: [remove|cancel|delete] [filter on|] headline [filters|criteria|keywords|search|] [and|,] [find|show] [me|] [articles|] [relating to|about|linked to|] l_9 [instead|]. In India, two participants use the pattern [kindly|] change [the|] list [that|to] contain l_9 related [articles|] [instead of having one of these l_4, l_5 or l_6 in the headline|ignoring the previous headlines I mentioned]. The adopted patterns in the United States are like please include only articles about l_9, but remove the filter about headlines containing certain words or remove my earlier headline requests and change the list to include only articles about l_9. There are other good patterns in these countries. For instance, repeat last search with headline criteria removed and include articles that mention l_9 in Australia, can you please limit news to be about l_9 and remove headline conditions and reset headline filter, new filter: contains l_9 in Canada. In the United Kingdom and India, a few participant use [with|] any headline to represent the headline conditions should be removed.

Task 11b Update($f_1 : (l_4 \vee l_7 \vee l_6), \neg(f_1 : (l_4 \vee l_7 \vee l_6)) \wedge l_9$): update part of filters to negation of filters and add a new filter

The misunderstanding utterances of Task 11b are mainly adopted from Australia and Canada. The patterns could be like please change the list to include articles about l_9 and remove articles with headlines that include l_4, l_5 or l_6, give me a list of articles that are about l_9 which instead of a headline that contains at least one of l_4, l_5 or l_6, please gather all news that are about l_9 that do not have l_4, l_5 or l_6 in the headline from Australia, and please find articles about l_9 instead of articles with l_4, l_5 or l_6 in the headline from Canada. Most participants from India and the United States misunderstand l_9 as one of the conditions for field, representing the operation Update(f_1:($l_4 \vee l_7 \vee l_6$), $\neg$(f_1:($l_4 \vee l_7 \vee l_6$))$\wedge$ f_1:(l_9)). Thus, the patterns of Task 11b do not cover this kind of misunderstanding utterances.

Task 12 Update(Q_{11}, l_{10}): replace entire query with new filter

The country that has the most utterances which emphasize the new search is Canada with the pattern [please|] [remove|discard|cancel|ignore|delete|forget about] [all|] [the|my|] [previous|past|] [filter|command|search|requests] [criteria|] [I gave you before|] [and|,] [only|just|] [search|find|show|get] [me|for|] [news about|] l_{10} [only|]. In other countries except Canada, the widely used pattern is [now|only|] [show|get|find|include|list|give] [me|] [all|different|a list of|] articles [about|

that have|related to|for|on] l_{10}. However, in the results of the evaluation, it shows that the utterances without now, only, different will cause the misidentification with the intent of Task 2 (Create Query) since there are overlapping patterns between two tasks. The patterns in Task 12 should be distinguished from the patterns in Task 2, even if the differences are minor. [start|conduct|] [a|] [new|fresh] search ... occurs in the other four countries except the United States, while it exists [delete the earlier list|disregard all previous searches|do not consider anything I said earlier] in the United States. In Canada and the United Kingdom, a few participants specify a new list rather than a new search as [please|] [create|show me|give] [a|] new list of articles [for|about|on] l_{10}. In Australia, Canada and the United Kingdom, a common expression is [please|] ignore [all|my|] previous instructions.

Task 13 Create($\neg$ l_1): add negation of filter to query

Except India, the other four countries have the most popular pattern as [exclude|without|take out|take away|exclude|eliminate|remove|disregard] [any|all|] [news|anything|results] [about|related to|on] l_1 [from|] [the|last|this|] [search|query|list|]. The last part of the pattern [from|] [the|last|this|] [search|query|list|] mainly occurs in Canada and the United Kingdom. In addition to the common prepositions [about|related to|on] to connect news and l_1 in these countries, participants from Canada use other prepositions such as [relating to|that talk bout|that mention], while participants from the United Kingdom have other preferences of prepositions like [including|which reference|which cover|relating to]. In India, participants prefer the pattern [show|give] me [the ones|only those articles|the list as I asked before but] that [do|are|is] not [refer to|about] l_1, followed by [just|] news that are not about l_1 and [hide|remove|exclude] articles [about|related to] l_1 [from the list|]. In the United States, another prevailing pattern is I [want|would like] [to see|] [news|list to contain articles about the previous topic] [that|] [are|] not about l_1. In Australia, Canada and the United Kingdom, a few participants use the pattern do not show [me|] any articles about l_1. However, this pattern overlaps the pattern of Task 7 (Update(l_5, ?)) such that it will lead to the misidentification of different intents. There is no doubt that such pattern should be removed from patterns of either Task 7 or Task 13.

Chapter 7

Conclusion

In this chapter, we recapitulate the whole thesis, followed by summarizing the results of the study. Finally, we present the limitations of this thesis and the future work ahead to cope with these unsolved problems.

The thesis's goal is to enable seekers to pose complex queries by clarifying cumulatively in a conversational search dialogue. The proposed solution is to build a natural and functional conversational search system that is able to rewrite the existing queries recursively. First, a study is undertaken to investigate how the system and seekers from different countries (Australia, Canada, India, the United Kingdom, and the United States) engage in different topic-oriented search sessions (argument, book, news, and trip). Seeker interaction and utterance intent in distinct scenarios or different countries are analyzed and characterized in the results.

In what follows, we present a summary of the results as well as the findings of the study.

Utterances are mainly commands: In the 2919 natural language utterances that labeled as "good", 60.6% of utterances are commands, starting with terms like "*show me*", "*find me*", "*give me*" "*search for*", "*filter*", "*include*", "*remove*", "*exclude*", "*do not include*", "*do not show me*". The number of questions and statements are close with 571 and 578 respectively. The utterances in the form of question start with "*can you show me*", "*can you find*", "*can you include*" "*can I see*", "*can I have*", "*what did I*", "*what was*", while statements begin with "*I am looking for*", "*I want*", "*I mean*", "*I would like to*", "*I did not mean*", "*I do not want*". In particular, the utterances in Task 7 and Task 8 are mainly statements rather than commands since seekers specify that they did not mean certain value to point out the system's error and then interpret the real value they actually meant. In Task 10, there are the most utterances as questions as they are asking whether the system is able to go back to or repeat the previous search.

For most tasks with different formulation intents, the basic structures of the popular patterns will not be affected by the scenarios: Two exceptions are in Task 6 and Task 13. In Task 6, as the selected fields are different, in the argument and book scenarios, participants prefer specifying the fields [source|genre] before the keywords, which is opposite in the news and trip scenarios with the fields [headline|hotel]. In Task 13, more participants indicate the created negation value before the collection type in the trip scenario, which is different from other scenarios. In the trip scenario, there are more unique patterns. For instance, it exists unique prepositions to connect the collection type and filters such as [to|for|about]. As the filters l_2 and l_3 already include a preposition "*by*", most utterances including these filters omit the preposition. In the argument and book scenarios, there are more similar patterns.

In half of the designed tasks, the structures of the most commonly used patterns in different countries are similar: Additionally, these tasks have more qualified utterances and fewer participants misunderstand. In each task, Canada and the United States have more unique patterns. Considering all tasks, the popular patterns in the other four countries except Canada are the most similar. To compare the similarity between the two countries, Canada and the United Kingdom have more similar patterns.

The same word in different positions of the utterances is likely to lead to different interpretations: For instance, for some restricted words such as [just|only], the interpretations in the utterance *can you only show me trips that are by ship* and in the utterance *can you show me trips that are only by ship* are different when also with reference to the previous results (detailed in Page 48). Another similar example is the word also in Page 51.

The usage of the pronouns is helpful for context: For an "intelligent" conversational search system that can reference the context of the whole search dialogue, the usage of pronouns is to remind the system that the search is related to the already mentioned concepts such that it needs to go back to the previous search to figure out what the pronouns refer to. In other words, it is found that using pronouns is helpful to solve the ambiguity in some tasks such as Task 3 (in Page 48) and Task 4 (in Page 49).

Some minor differences can be a good sign for distinguishing by different intents: Through the evaluation, the results reveal that some words like instead, only, just, now, different is a good sign to distinguish from starting a search session and restarting a completely new search, even if the rest of the utterances are identical (in Page 56).

However, there are a few restrictions in the study and the front-end of the prototype that remain to be solved for future work. In the study, participants are recruited through an online crowdsourcing marketplace and virtually conduct the study. We collect the participants' answers and subjectively analyze their utterances without confirming the intents of their utterances with them again. Thus, it is hard to know whether the participants misunderstand the tasks or formulate the utterances inappropriately by accident, especially for those ambiguous utterances. In the end, the ambiguous or misunderstanding utterances are not considered in the organized query reformulation patterns. But in general, crowdsourcing is an optimal way to take the least time cost to get as many answers as possible in the early stages. When the prototype is fully built in the future, it is worth implementing a user study that observes how seekers use the prototype. The user study is allowed to interact with the participants for the experimenters and receive feedback from the participants in time. We can recruit the participants who perform well in the early study to conduct the user study again since they have good knowledge of conversational search and also have experience in our study.

On the one hand, for the front-end of the prototype, it is not allowed to exist any punctuations in the sample utterances of each intent in the interaction model. Nevertheless, the natural language utterances are collected in a text-based chat interface such that punctuations are inevitable, especially when participants present complex information needs with multiple requirements, they prefer using punctuations to connect several sentences. The punctuations from the natural language utterances are discarded in the sample utterances of each intent. Unfortunately, such a way by removing punctuations is likely to not only lead to the ambiguity but also increase the difficulty in analyzing and understanding of annotations for Alexa. A example can be in Task 6 with the operation Create($f_1 : (l_4 \wedge l_5 \wedge l_6)$): an annotation *show the source of these arguments such as from BestReasons, WhatsUp or WikiDiscussions* corresponds to the pattern show the {field} of these {collection} such as from {filterOne} {filterTwo} or {filterThree} without punctuations. This annotation is failed to have the expected intent CreatePartIntent. In one of the Alexa's interpretations for this annotation, the actual value of the slot type {filterOne} is mapped with *Best Reasons what's*, while {filterTwo} is mapped with *up* and {filterThree} is mapped with *wikidiscussions*. Although the lists of the possible values of these three slot types are identical. The problem mainly occurs in the book scenario since filters' values are more complicated than the other three scenarios. Besides Task 6, these misinterpretations commonly appear in other intents with multiple slot types such as UpdateFilteralIntent (Task 9) and UpdatePartFieldIntent (Task 11). That is to say, if the sample utterances comprise multiple slot types without any words or punctuations as separations, it is hard for Alexa to interpret the corresponding annotations correctly. Hence, it could lead to a study that aims to explore how seekers say a series of filters consecutively without the help of textual punctuations in a speech-based search interface and how the textual requests look like, which are transformed by the system from the speech inputs. It is interesting to see if participants perform differently when they need to indicate complex information needs in the search interfaces with different mediums. Another way to ensure that multiple slot types are assigned with the exact values is to differ from the lists of different slot types' possible values. But this way is more applicable to those slot types that have little connections.

On the other hand, since Amazon Alexa restricts the generated interaction model's file size, the organized query reformulation patterns can not be applied in the sample utterances of intents. Alternatively, the collected natural language utterances are regarded as the sample utterances by replacing the specific values with the abstracted slot types. Such a way is at the expense of the diversity and generalizability of sample utterances, which is also reflected by cross-evaluation results (Section 5.3). In cross-evaluation, a large number of annotations fail to be assigned with the expected intents since Alexa is strict with each difference between the human annotations and the sample utterances. A word difference such as "*me*", "*all*" or "*a*" can result in the misidentifications of the intents, which is shown in Page 42 and in Page 44. In fact, these misidentified annotations can be tackled by using the summarized patterns as sample utterances since the summarized patterns can extend the possibilities of the sample utterances and not only limited to the collected natural language utterances. A typical example can be seen in Page 45. While developing the interaction model, other limitations of the current agent Alexa can be the definition of the slot types. The slot type AMAZON.SearchQuery that captures less-predictable inputs by users can only be used once in a sample utterance without any other custom slot types. As a result, developers must define each custom slot type with a list consisting of all possible values manually. In addition, the synonyms of certain values are determined if needed. If the developers overlook some values, the annotations are likely to fail to have the expected intent. For instance, the synonyms of the field hotel can be hotel accommodation or accommodation. The problems mentioned above are worth solving further to develop a more intelligent and powerful search application.

Finally, due to the time constraints, the prototype's back-end is still in the conceptual stage. A follow-up study can be targeted towards bringing the proposed query rewriting layer to practice. Also, a foreseeable challenge is to not only predict the ambiguity of human annotations but also minimize and convert into unambiguous queries.

Appendix A

Patterns

Task	Scenario	Country	Template	Frequency
2	news	AU	[please|] [show|get|gather|give] [me|] [all|a list of|] [recent|latest|] news [about|on] l_1.	9
			[latest|list of|] news [about|on|of] l_1.	4
			[list|] l_1 news.	3
		CA	[please|] [find|get|show|tell|give|create] [me|] [all|some|a list of|] news [on|about|related to|relating to] l_1.	16
			can you [please|] [get me|retrieve] [all|] news [on|about] l_1?	3
			give me l_1 news.	1
		GB	[please|] [find|get|show|filter|give] [me|] [all|a list of|] news [on|concerning|about|related to|regarding|which cover] l_1.	12
			[can|could] you [find|show|get] me news [on|about] l_1?	3
			[all|] news [about|on] l_1.	2
			find all l_1 news.	1
		IN	[please|hey,|] [get|show|tell|fetch|list] [me|] [all|some|] [latest|] news [on|about|related to] l_1.	14
			[all|a list of|] news [about|on] l_1.	4
		US	[find|get|tell|give] [me|] [all|] [latest|] news [on|about] l_1.	7
			[please|] [find|give|provide|search for|retrieve] [me|] [latest|any|all|] l_1 news [you can for me|].	5
			news [about|on] l_1.	2
	argument		[please|] [find|tell|search|provide|give|get|list|obtain|fetch|show|] [me|] [all|some|a list of|] [arguments|reasons] [on|about|for|regarding|related to|that are for and against] l_1.	28
			[can|could] you [please|] [give|get|list|tell|find|obtain] [me|] [some|a list of|all|] [good|popular|] arguments [for and against|on|about|regarding] l_1?	11
			can I [see|have|get] [a list of|all|] arguments [on|about|concerning|for] [the advantages and disadvantages of|] l_1?	5
			I am looking for the pros and cons [of|on] l_1.	2
			what do you think about l_1?	1
	book		[hi|] [please|] [get|find|fetch|show|recommend|list|give|] [me|] [a list of|all|some|those|] books [on|about|related to|which describe about] l_1.	30
			[hi|] [can|could] you [please|] [find|get|list|look up|search|show|send] [me|] [a list of|some|all|] books [on|about] l_1?	16
			can I [have|see] [a list of|all|] books [about|on] l_1?	7
			I [am looking for|want|would like] [all|] books about l_1.	4
	trip		[please|] [show|find|get|give|list|] [me|] [a list of|all|] [options for|] trips to l_1.	26
			[can|could] you [please|] [show|make|give|open up|find|get] [me|] [a list of|all|] [available|different options on|any information about|] trips [for|of|to] l_1?	16
			[hello|] I [want|need|would like|am looking for] [to see|you to get|] [a list of|all|] trips to l_1.	8
			[please|] can I [see|have] [a list of|all] trips [to|headed for] l_1?	6
			hello I like to plan a trip for l_1, would you help me to do that?	1
			what trips are available to l_1?	1
3	news	AU	show me [only|] the ones about l_2.	2
			show more details about its l_2.	1
			I want the ones that talk about l_2.	1
		CA	[filter|keep] articles [to|] [containing|mention only|related to|about] l_2.	4
			I want it to contain only news about l_2.	1
			please filter articles to ones that contain l_2.	1
			narrow the list down to articles about l_2.	1
			reduce list to only show l_2 articles.	1
			filter out articles that do not mention l_2.	1
		GB	show [me|] [just|only] [ones|those] articles [that contain information|] about l_2.	2
			remove all articles which do not cover the subject of l_2.	1
			limit to l_2 news.	1
		IN	all those related only to l_2.	1
			show me only ones about l_2.	1

<table>
<tr><td>3</td><td></td><td></td><td>could you only show me results that are about l_2?</td><td>1</td></tr>
<tr><td></td><td></td><td>US</td><td>can you just show me the ones that discuss l_2?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>show me only those about l_2.</td><td>1</td></tr>
<tr><td></td><td>argument</td><td></td><td>can you [filter|shorten|trim down|narrow down] [these|that|list|arguments|] [to|] [ones|] [only|just|] [with|about|related to|regarding|] l_2?</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td>[show|give] [me|] [only|just] [the|] ones [that are|] about l_2.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>which [of|] these [include|relate to|pertain to] l_2?</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>[only|] show [me|] results that [also mention|include] l_2.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>filter [these|this list] [with|for] l_2 [only|].</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>give me arguments that focus on l_2 from them only.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>remove all arguments that do not related to l_2.</td><td>1</td></tr>
<tr><td></td><td>book</td><td></td><td>[narrow|filter|limit] [down|] [list|search] to [books|those|] [about|pertaining to|by|for|] l_2.</td><td>7</td></tr>
<tr><td></td><td></td><td></td><td>[just|] [show me|pick] the ones [that are|] [about|on] l_2.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>[please|] remove [all|] books [that are|] not about l_2.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>can you remove books that are not related to l_2?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please show me all books about with l_2 in them.</td><td>1</td></tr>
<tr><td></td><td>trip</td><td></td><td>[just|] show [me|] [the|] [ones|those] [trips|] [that are|] l_2.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>which ones are [trip|] l_2?</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>narrow list to the ones l_2.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>filter with only trips l_2.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>I want to have a travel to there l_2.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>remove all trips that are not l_2.</td><td>1</td></tr>
<tr><td>4</td><td>news</td><td>AU</td><td>l_2 or l_3 articles.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please gather all news that are about l_3 that would be an alternative to l_2.</td><td>1</td></tr>
<tr><td></td><td>news</td><td>CA</td><td>alter previous filter and keep articles containing the word l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>can I just see the articles that only contain news about l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>can you get me more news on l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td>IN</td><td>include ones about l_3 as well.</td><td>1</td></tr>
<tr><td></td><td></td><td>US</td><td>can you also include ones that discuss l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>I want to see news about either l_2 or l_3.</td><td>1</td></tr>
<tr><td></td><td>argument</td><td></td><td>[I am looking for arguments about|filter with] either l_2 or l_3.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>can you give me the same arguments but add those including l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>can you also show me ones about l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>I would like a list of arguments related to l_2 as well as those arguments related to l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>are there any arguments about l_2 as an alternative to l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>show me the ones about l_3 too.</td><td>1</td></tr>
<tr><td></td><td>book</td><td></td><td>can you [find|search] books about l_3 or l_2 [please|]?</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>could you also include ones about l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>could you also suggest me something on the same topic but for $_3$?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>could you change it to l_3 or l_2?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>add l_3 to filter as alternative.</td><td>1</td></tr>
<tr><td></td><td>trip</td><td></td><td>[now|] [please|] [find|show] [me|] [all|] [trips|] [that are|] [either|] l_2 or l_3.</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>[or|another option is] l_3.</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td>can you include the ones l_3 too?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>show me trips l_2, if l_2 trips not available then I would like to select trips l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>if there are no choice to l_2 transport , will alternatively show me trip l_3.</td><td>1</td></tr>
<tr><td>5</td><td>news</td><td>AU</td><td>show me [all|more] news [not just|other than] [the|] ones about l_2 and l_3.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>can you remove the l_2 and l_3 filters?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[now|] [please|] [show|tell|give] [me|] [more|all|a list of|] articles [that are not|except|that do not|not] [about|discuss|on|related to|] l_2 or l_3.</td><td>7</td></tr>
<tr><td></td><td></td><td>CA</td><td>can you please get me all articles not just limited to l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>can I see more articles that show information other than l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please remove the word filters l_2 and l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[remove|exclude|take away] [articles|] [on|about|] l_2 and l_3 [from|] [the|original|] [list|] [but include everything else|].</td><td>3</td></tr>
<tr><td></td><td></td><td>GB</td><td>can you change the articles to show information other than l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please remove filters on l_2 and l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>show me all articles, not just l_2 and l_3 ones.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[please|] [get|find|show] [me|] [all|] articles [that are|] not [about|related to] l_2 or l_3.</td><td>3</td></tr>
<tr><td></td><td></td><td>IN</td><td>[get|show|] [me|] [all|] news [apart from|other than|not just] l_2 [or|and] l_3.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>can you list those news other than related to l_2 or l_3?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[now|] [please|] [show|tell] me [all|] articles [that are|] not [related to|on] l_2 or l_3.</td><td>3</td></tr>
<tr><td></td><td></td><td>US</td><td>[actually|] also [add|get] articles [that are|] not [about|related to] l_2 or l_3.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>show articles other than those about l_2 or l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>include articles that do not include l_2 or l_3 too.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>I [only|] [want|would like|need] [to see|] news [that are|] not about l_2 or l_3.</td><td>4</td></tr>
<tr><td></td><td>argument</td><td></td><td>[please|] add [back|] [all|any|] [original|relevant|] [arguments|ones] [that are|] not [about|related to] l_2 or l_3 [to the list|].</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td>[other than|besides|including] l_2 and l_3, [what other factors are at play|what are other advantages and disadvantages|are there other reasons]?</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>forget about only including l_2 and l_3.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[now|] [please|] [show|tell|find out|find|] [me|] [more|all|other|] arguments [that are|] not about l_2 or l_3.</td><td>10</td></tr>
<tr><td></td><td>book</td><td></td><td>[please|] [expand the list to|] [add|include] [all|] books that are not about l_2 or l_3.</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td>remove l_2 and l_3 from filter.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[please|] [only|] [keep|find|make|change] [a list of|the list to|] books [in|from|] [the|] [original|] [list|] [that|] are not about l_2 or l_3.</td><td>5</td></tr>
<tr><td></td><td>trip</td><td></td><td>can you [update the list to|] [show|provide] [me|] [all|] trips [not just|in addition to|other than] [the ones|] l_2 and l_3?</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>show [me|] [more|] [trips|full list] [by all transport|] not just [those|] l_2 [or|and] l_3.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>remove filters [by|for] l_2 and l_3 [and show me all the trips|].</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>besides l_2 or l_3, what other option do you have, in order to get there?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>[now|] [please|] show [me|] [all|] [other|] trips [to that place|] [that are|] not l_2 or l_3.</td><td>12</td></tr>
</table>

6	news	AU	[please|] [only|] [show|give|include] [me|] [only|] [a list of|all|] articles [which|that|] [have|mention|include|references|contain|with] l_4, l_5 or l_6 in the headline.	9
			[show me|] articles with headlines [related to|including] l_4, l_5 or l_6.	2
		CA	[please|] [only|] [show|find|give|modify the list to show|include] [me|] [just|only|] news [that contain|with|that have] l_4, l_5 or l_6 in the headline.	10
			[please|] [show|update|narrow the list to] [me|] [only|] news [with headlines containing|with headlines mentioning|only containing headlines with] l_4, l_5 or l_6.	3
		GB	[now|] [please|] [only|] [get|show|narrow it down to] [me|] [only|] news [that contain|with|that include] [either|] l_4, l_5 or l_6 in the headline.	8
			[now|] show [me|] [just those|] news [where the headline includes at least one of the following terms:|with headlines:] l_4, l_5 or l_6.	2
			remove articles where the headline does not contain l_4, l_5 or l_6.	1
		IN	[get|show] [me|] [just|only those|all|] news [that contain|with|which have|in which] [either|] l_4, l_5 or l_6 in the headline.	9
			news [that|with] headline [contains one of|for|having the terms like|with] l_4, l_5 or l_6.	4
		US	[only|] [show|find|include] [me|] news [with|that have] l_4, l_5 or l_6 in the headline.	6
			[please|] [show|include] [me|] [all|] news [with headlines related to|that have the headline of either] l_4, l_5 or l_6.	2
	argument		[only|] [search|select|show|include|give] [me|] [results|arguments|] [that|] [are|have|] [from|] [a|the|] [following|] sources [that include|for|like|are|from|of|] l_4, l_5 or l_6.	17
			can you [please|] [just|only|] [show|find|return] [me|] arguments [that|] [come|have|] [from|] [a|the|these] [following|] sources [web page|] [of|on|] l_4, l_5 or l_6?	5
			please remove from the list any arguments where the source is not one of: l_4, l_5 or l_6.	1
	book		[only|] [get|find|give|make|bring|show|list] [me|] [a list of|] books [that|] [are|] [either|] [in|from|with|where|with|] [one|] [of|] [the|these|a|] [following|] genre [has|belonging to|of|] l_4, l_5 or l_6 [from the list|in the name|].	13
			[can|could] you [only|] [get|find|show|select|filter] [me|] [a list of|all|the|those|] books [that are|] [within|in|from|to] the genre [of|containing|] l_4, l_5 or l_6?	6
			[please|] remove books that [are not|do not have] l_4, l_5 or l_6 [in the|] genre [name|].	4
	trip		[please|] [just|] [show|give|] [me|] [all|only|] [trips|] [involving|to|that include|for|headed towards|with|that are booked at] l_4, l_5 or l_6 [as the|] hotels.	12
			show [me|] [all|] [ones|trips] [with|where I stayed] [in|at] [the following|one of these] hotels: l_4, l_5 or l_6.	4
			in your list, which ones include accommodation in l_4, l_5 or l_6 hotels?	1
			remove trips where hotel is not l_4, l_5 or l_6.	1
7	news	AU	l_5 [is|does] not [stand for|mean|] {?}.	6
			{?} is not correct.	1
			please correct results, l_5 does not refer to {?}.	1
			I am not asking for {?}.	1
		CA	l_5 [is|does|] not [equal|refer] [to|] {?}. [please correct that for me.|]	3
			I [do|did] not mean {?} [for l_5|].	2
			I did not mean l_5 as {?}.	1
		GB	I [do|did] not mean {?} [when I said l_5|].	4
			I did not [intend for|mean] l_5 [to refer to|as] {?}.	2
			l_5 is not {?}. [please try again.|]	2
		IN	l_5 [does not mean|not refers to|that I meant is not] {?}. [please revise|].	4
			I [did|was|do] not [mean|referring to|want for] {?} [for l_5|].	4
			just correct l_5, does not mean {?} so correct it.	1
			I did not mean l_5 as {?}.	1
		US	l_5 [does|is] not [stand for|mean|] {?}.	6
			I did not mean {?} [by l_5|].	2
			{?} is incorrect.	1
			the interpretation you had for l_5 as {?} is not what I actually meant.	1
	argument		[you misunderstood me as|] I am not [looking for|talking about] {?}.	6
			[no,|] I did not mean [results from|] l_5 [as|] {?}.	3
			l_5 is [a source,|] not {?}.	3
			{?} is incorrect.	1
	book		I [did not mean|am not looking for] {?} [when I said l_5|by l_5|].	14
			[genre|] l_5 does not [stand for|mean] {?} [in this context|].	8
			l_5 [is|] not {?}.	5
			{?} is [incorrect|not what I meant].	2
			I did not mean l_5 [to translate to|as] {?}.	2
	trip		[actually|] I [did not mean|was not referring to|was not looking for] {?}.	8
			[note that|] l_5 [is|was|does] not [supposed to be|mean|referring to|] {?}.	8
			I [am|did] not [mention that|mean] l_5 [as|to be] {?}.	3
			I [did not meant|was not referring to] {?} when I [asked for|said] l_5.	2
			{?} is incorrect.	1
8	news	AU	I [meant|was referring to] [to say it is|] l_7.	6
			l_5 [means|stands for|should refer to] l_7.	4
		CA	I [meant|mean|was looking for|was referring to] [l_5 to search for|] l_7.	8
			l_5 [stands for|means|is equal to] l_7 [and not {?}. Adjust accordingly.|]	5
		GB	I [actually|] [meant|mean|want] l_7. [search for it.|]	7
			l_5 [is|means|may be interpreted as] l_7.	4
		IN	I [meant|mean|want] [l_5|that|] [as|for|] l_7 [in mind|].	7
			[it is|] l_7.	5
		US	I [actually|] [meant|mean] l_7 [by l_5|].	5
			l_5 [actually|] [means|stands for|is] l_7.	5
			l_7 [articles|].	4
			l_5 is an abbreviation for l_7.	1
			What I had in mind initially was l_7.	1
	argument		I [am|was|actually|] [meant|want|need|would like|looking for|talking about|thinking of|] [to|] [hear|see|] [arguments|results|] [sources|] [from|] [l_5|] l_7.	32
	book		I [actually|was|] [meant|talking about|would like|need|talking about|thinking of] [to see|] [books|] [genre|] [l_5 as|] l_7 [when I said|for|] [l_5|].	27

<table>
<tr><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td></td><td></td><td></td><td>[the|] [genre|] l_5 [means|stands for|refers to|denotes] l_7.</td><td>9</td></tr>
<tr><td></td><td></td><td></td><td>could you change {?} to l_7?</td><td>1</td></tr>
<tr><td></td><td>trip</td><td></td><td>I [actually|was|] [meant|wanted|looking for|thinking|talking about] [to|] [say|include trips to|stay in] [l_5|] [to be|equal to|] l_7 [hotel|] [by l_5|].</td><td>31</td></tr>
<tr><td></td><td></td><td></td><td>change filter from {?} to l_7.</td><td>1</td></tr>
<tr><td>9</td><td>news</td><td>AU</td><td>[please|] [search for|show|gather|give] [me|] [all|a list of|] articles [that are|] [relating to|on|about] l_8 [instead of|not] l_1 [with|] [the same|these|] [headline|] [filters|].</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td>[using the same filters,look for|same query but for] l_8 [articles|] [instead of|not] l_1.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>[replace|change] l_1 [with|to] l_8 [in the previous headline search|].</td><td>2</td></tr>
<tr><td></td><td></td><td>CA</td><td>[now|] [please|] [show|search|find|give] [me|] [some|a list of|] news [about|] l_8 [instead of|remove articles] [about|] l_1 [with|keep|] [those|the same|] [headlines|] [filters|].</td><td>8</td></tr>
<tr><td></td><td></td><td></td><td>I [want|would like] to see news [mentioning|about] l_8 [and|] [not about|instead of] l_1.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>instead of articles [about|] l_1 [that match the criteria I gave you|], [please|] [search|show me] news about l_8 [that match the criteria|].</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>in addition to the previously mentioned keywords in the title, change the articles to include l_8 instead of l_1.</td><td>1</td></tr>
<tr><td></td><td></td><td>GB</td><td>[now|] [please|] [adjust that query to|] [get|show|search for|do a new search on|search] [me|] news [containing|about] l_8 [and|,|] [instead of|not|remove results about] l_1.</td><td>8</td></tr>
<tr><td></td><td></td><td></td><td>can you find me articles with those initials in the title to do with l_8 rather than l_1?</td><td>1</td></tr>
<tr><td></td><td></td><td>IN</td><td>[now|] [get|show|add] [me|] [news|results] [including|about|related to|that meet the criteria above for] l_8 [and|,|] [instead of|rather than|excluding] l_1.</td><td>9</td></tr>
<tr><td></td><td></td><td>US</td><td>I [want|would like|need] [to see|] news [mentioning|about|on] l_8 [not|instead of] [on|]l_1.</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>[please|] [show|looking for|give] [me|] news [on|about] l_8 [and|,|] [instead of|not|rather than] l_1 [,they should meet the previous criteria|].</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>I want the articles about l_1 to be about l_8 now with the same headlines.</td><td>1</td></tr>
<tr><td></td><td>argument</td><td></td><td>[please|] [include|search for|list|show|add|tell|give|] [me|] [all|] [arguments|results|pros and cons|] [about|that include|based on|for|that focus on|related to] l_8, [and|but|] [not|rather than|instead of|as opposed to|delete] l_1 [from|] [the same|these|] [sources|].</td><td>15</td></tr>
<tr><td></td><td></td><td></td><td>please generate a list of arguments for l_8 instead of l_1 using the same criteria.</td><td>1</td></tr>
<tr><td></td><td>book</td><td></td><td>[only|] [get|find|fetch|show|list] [me|] [different|a list of|] [books|] [that|] [are|] [about|containing|have] l_8, [instead of|not|without containing|rather than] l_1 [in|from|on|by|] [the same|those|] [genres|] [that I mentioned|as before|].</td><td>18</td></tr>
<tr><td></td><td>trip</td><td></td><td>[now|] [please|] [show|give] [me|] [all|] trips to l_8 [not|instead of|rather than] l_1.</td><td>10</td></tr>
<tr><td></td><td></td><td></td><td>[I want to|] [change|replace] [trip|] [destination|] [from|] l_1 to l_8 [keep|with|] [the|] [same hotels|] [destinations|].</td><td>10</td></tr>
<tr><td></td><td></td><td></td><td>could you show me the same criteria but for l_8, not l_1?</td><td>1</td></tr>
<tr><td>10</td><td>news</td><td>AU</td><td>[please|] [repeat|show|remind me of] [me|] [my|the|] [previous|last|current|past] [query|message|search|inputs] [criteria|filters|].</td><td>8</td></tr>
<tr><td></td><td></td><td></td><td>what did I [just ask to search for|say|tell you to do before]?</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>what was my last [query|request]?</td><td>2</td></tr>
<tr><td></td><td></td><td>CA</td><td>[please|] [show me|show|remind me of] [my|] [previous|] [search|requests] [criteria|].</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td>can you [please|] repeat [my|the] [previous|last] [information I provided|request]?</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>can you please tell me what [the latest search parameters were|I had previously requested from you]?</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>what did I [previously ask you to do|search for]?</td><td>2</td></tr>
<tr><td></td><td></td><td>GB</td><td>[list|repeat|show|tell me my|remind me of] [previous|current|] search [details|entries|].</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>what was my [last|previous] search [criteria|parameters|]?</td><td>4</td></tr>
<tr><td></td><td></td><td>IN</td><td>[show me|display|repeat] [my|all|] [previous|last|] [command|query|search|filters].</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td>what was my previous [request|message]? [I forgot which news I asked for|].</td><td>2</td></tr>
<tr><td></td><td></td><td>US</td><td>what was [my|the] [last|previous] [request|search] [query|]?</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>[please|] [go back to|show me|repeat] my previous [search|request].</td><td>3</td></tr>
<tr><td></td><td>argument</td><td></td><td>[please|] [show|tell|repeat|confirm|find] [to|] [me|] [about|] my [previous|earlier|] [query|request|argument criteria|messages|question].</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td>[sorry, I lost track of where we were,|] can you [show|repeat] [me|] what I [previously|] [have asked before|told you about arguments|asked you to do|have searched for]?</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td>can you [please|] [show|remind|repeat|tell|recap] [me|] [filters applied to list|parameters of search|your current criteria|of my previous commands]?</td><td>4</td></tr>
<tr><td></td><td>book</td><td></td><td>[please|] [tell|show|give|remind] [me|] [of|] [my|] [last|previous|] [message|search|command] [history|parameters|keywords|].</td><td>9</td></tr>
<tr><td></td><td></td><td></td><td>[I forgot what I asked you previously,|] can you [tell|remind|show] [me|] what [I asked you to find earlier|was my last search about|this list is based on|filters have you currently applied|I previously asked you|searched for previously|I asked for earlier]?</td><td>7</td></tr>
<tr><td></td><td>trip</td><td></td><td>[please|] [show|list|tell] [me|] [my|] [previous|original|last|] [search|filters|requests|query] [details|criteria|].</td><td>9</td></tr>
<tr><td></td><td></td><td></td><td>[sorry, I got confused.|] [can|could] you [please|] [tell|remind|repeat|show] [me|] what [I previously searched for|information I asked|my previous inquiries were|my last trip search was|my previous question was|I asked you previously]?</td><td>7</td></tr>
<tr><td>11</td><td>news</td><td>AU</td><td>remove [current|] headline [filters|criteria], show me articles relating to l_9 [instead|].</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>repeat last search with headline criteria removed and include articles that mention l_9.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>give me a list of articles that are about l_9 which instead of a headline that contains at least one of l_4, l_5 or l_6.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please change the list to include articles about l_9. Remove articles with headlines that include l_4, l_5 or l_6.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please gather all news that are about l_9 that do not have l_4, l_5 or l_6 in the headline.</td><td>1</td></tr>
<tr><td></td><td></td><td>CA</td><td>[cancel|delete] headline [filter|keywords], find [articles|] [about|] l_9.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>can you please limit news to be about l_9 and remove headline conditions?</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>reset headline filter, new filter: contains l_9.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>please find articles about l_9 instead of articles with l_4, l_5 or l_6 [in the|as] headline.</td><td>1</td></tr>
<tr><td></td><td></td><td>GB</td><td>[remove|cancel] [headline search|filter on headline], [find|show] [articles|] about l_9.</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td>[please|] search for articles [on|about] l_9 and [with|] any headline.</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td>cancel search filter for articles containing l_4, l_5 or l_6 in the headline. Can you show me articles that mention l_9?</td><td>1</td></tr>
<tr><td></td><td></td><td>IN</td><td>[kindly|] change [the|] list [that|to] contains l_9 related [articles|] [instead of having one of these l_4, l_5 or l_6 in the headline|ignoring the previous headlines I mentioned].</td><td>2</td></tr>
</table>

11			show articles from last query about l_9 and [with|] any headline.	1
		US	include only articles about l_9, remove filter about headlines containing certain words.	1
			remove my earlier headline requests, change the list to include only articles about l_9.	1
	argument		[please|] [include|find|give|show|limit] [me|] [only|] [arguments|pros and cons|results] [that are|] [about|related to|relating to] l_9 [and|] [pay no attention to|arguments can come from any|from any|delete|get rid of any] sources [instructions|requirements|].	7
			I want arguments about my previous query about l_9, I do not care about source.	1
			[show|tell|] [me|] arguments [that|] [focus on|about|related to] l_9 [and|,|] [not from any source previously mentioned|without a source from before|not from previous source].	3
	book		[clear|stop|cancel|remove|delete] genre names [from|] [filter|], [filter by|only|search for|search|show me books about] l_9 [books|].	5
			I want the list to be only of those where genre name does not matter, but they have l_9.	1
			I [would like|need|want] [a list of|] books [on the same topic but|] [about|with|that have|on] l_9 [instead of one of|without] [those|] genre names [that I mentioned before].	4
	trip		[please|] [now|] [show|give|include] [me|] [trips|] [that|] [are|] [with|involve|about|include] l_9 [instead of|not|rather than] hotels.	8
			[remove|clear] hotel [filter|request|requirements], and [show me trips that are for|remove from list all trips that are not|include|change to|show|add] l_9.	7
			show me [all|] trips [that are|] [about|with] l_9, [the hotel doesn't matter|and remove hotel requirements|include all hotels].	3
			change list [,I just want trips that are about|to] l_9 [,I don't care|forget] about hotels.	2
12	news	AU	[now|] [show|list] [us|me|] articles [on|about|for] l_{10}.	8
			[clear my last search,|let us start from scratch.|start a new search.] I [want|am looking for] [to read|] [articles|] about l_{10}.	3
			[please|] [ignore|remove|disregard] [all|] previous [queries|articles found|instructions] and [now|] [show|give|gather] [me|] [a list of|all|] news [that are|] about l_{10}.	3
		CA	[please|] [remove|discard|cancel|ignore|delete|forget about] [all|] [my|] [previous|past|] [filter|command|search|requests] [criteria|] [I gave you before|] [and|,] [only|just|] [search|find|show|get] [me|for|] [news about|] l_{10} [only|].	8
			start a new [search|subscription] with [articles about|topic] l_{10}.	3
			[please|] [create|show me] [a|] new list of articles [for|about] l_{10}.	3
			please find news about l_{10} and ignore my previous instructions.	1
		GB	[only|] [get|show|find|include] [me|] articles [about|that have] l_{10}.	6
			[start|] [a|] new search [for|containing only] [articles about|] l_{10}.	3
			please ignore all previous instructions. can you show me articles about l_{10}?	1
			please give a new list of news on the subject of l_{10}.	1
		IN	[kindly|] [show|get|list|give] [me|] [all|] [news|list] [only|] [related to|about] l_{10}.	12
			forget [everything|all my previous requests] and [show|find] [me|] [all|] news [about|related to] l_{10}.	2
			fresh search on l_{10}.	1
		US	[now|] [only|just|] [give|show] [me|] [different|a list of|] [information|articles] about l_{10}.	9
			[delete the earlier list|disregard all previous searches|do not consider anything I said earlier] and [only|] [get|provide|change] [me|] [all|] [news|search results] [about|for] l_{10}.	3
	argument		[forget|disregard|ignore|remove] [all|] [my|] [previous|] [everything|search requirements|criteria|result parameters|commands|instructions], [and|] [now|] [please|] [give|show] [me|] [a list of|] [arguments|results|] [about|regarding|for|] l_{10} [now|].	8
			[I have a new request.|Let us start a brand new list.|] [now|this time|] [find|search for|show|start] [me|] [a|the|] [entirely|] [new|] [list of|] arguments [for|about] l_{10}.	8
	book		[do a new search|delete previous list|] [show|find|] [me|] [books|] [about|on] l_{10} [only|].	8
			[disregard|drop|forget|remove|clear] [all|] [current|previous|last|] [instructions|search] [filters|criteria|], [now|] [only|] [find|show|search for] [me|] [books|] [on|about|] l_{10}.	6
	trip		[now|] [please|] [show|find] [me|] [all|] [different|] trips to l_{10}.	8
			[start|] [a|] new [search|list|search criteria]. [search for|show] [me|] [all|] trips to l_{10}.	5
			forget [about|] [all|] [previous trips|last list|commands|everything] [from before|] [now|and] [show me|I want|list|open] [all|] [different|] trips to l_{10}.	5
13	news	AU	[exclude|disregard|without|take out|take away] [articles|results|any|] [about|related to|on|] l_1.	8
			[do not show any|change the list so that it does not include] articles about l_1.	2
		CA	[in your search,|] [please|] [exclude|eliminate|remove] [any|all|] [news|results] [from the list|] [relating to|related to|that talk about|about|that mention|on] l_1 [from the list|].	8
			[do not show me any|hide|without showing me] articles about l_1.	3
		GB	[please|] [now|] [remove|exclude|disregard] [all|any|] [articles|search results|] [including|about|which reference|which cover the subject of|relating to|] l_1 [from|] [the|last|this|] [search|query|].	11
			[do not show any articles about|minus] l_1.	2
		IN	[show|give] me [the ones|only those articles|the list as I asked before but] that [do|are|is] not [refer to|about] l_1.	5
			[just|] news that are not about l_1.	3
			[hide|remove|exclude] articles [about|related to] l_1 [from the list|].	3
			show me articles [except|not about] l_1.	2
		US	[do not include|exclude|remove] [all|any|] [anything|news|] [about|] l_1.	5
			I [want|would like] [to see|] [news|list to contain articles about the previous topic] [that|] [are|] not about l_1.	4
	argument		[ok|] [now|] [please|] [remove|filter out|take out|delete] [any|those|] [arguments|topics|entries|items|anything|] [that|] [are|] [about|related to|dealing with|mention|relate to|to do with] l_9 [from the list|].	20
			could you [please|] [remove|exclude|take away] [all|] [arguments|] [that are|] [about|for|] l_9 [from the results|] [and only include the arguments that I wanted from the previous instructions|]?	8
	book		[please|] [remove|exclude|filter out|discard] [all|] [books|] [that|] [are|] [with|related to|about|from|have|on|] l_9 [from|] [the list|search|] [of books|].	12
			[please|] [only|] [get|fetch|find|filter|adjust for|show|include] [me|] [only|] books [that|] [do|are|] [not|without] [have|about|from|] l_9.	12
	trip		[please|] [exclude|without|remove|eliminate|do not include|but not|forget|do not show me|discard] [any|all|] l_9 [from|] [trips|filters|list|].	20

<table>
<tr><td>[please|] [now|] [show|search|] [me|] [all|only|] [trips|just the ones|anything] [that|which|] [do|are|] [not|exclude|without] [involve|about|include|have|including|] l_9.</td><td>15</td></tr>
</table>

Table A.1: Collected reformulation patterns in four scenarios (argument, book, news and trip) occurred in the Chapter 6, where GB is the abbreviation of the United Kingdom.

Bibliography

Tawfiq Ammari, Jofish Kaye, Janice Tsai, and Frank Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction*, 26:1–28, 04 2019. doi: 10.1145/3311956. 1

Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, CHIIR '18, page 52–61, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349253. doi: 10.1145/3176349.3176380. URL `https://doi.org/10.1145/3176349.3176380`. 1

Guillaume Cabanac, Max Chevalier, Claude Chrisment, Christine Julien, Chantal Soulé-Dupuy, and Pascaline Tchienehom. *Web Information Retrieval*. 01 2008. doi: 10.4018/9781599047744.ch016. 1

J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). *SIGIR Forum*, 52(1):34–90, August 2018. ISSN 0163-5840. doi: 10.1145/3274784.3274788. URL `https://doi.org/10.1145/3274784.3274788`. 1

Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 223–232, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450323512. doi: 10.1145/2556195.2556217. URL `https://doi.org/10.1145/2556195.2556217`. 1

Norman Fraser, Dafydd Gibbon, Roger Moore, and Richard Winski. *Assessment of interactive systems.*, pages 564–615. Mouton de Gruyter, 1998. 1

Jiepu Jiang, Wei Jeng, and Daqing He. 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*, SIGIR '13, page 143–152, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320344. doi: 10.1145/2484028.2484092. URL `https://doi.org/10.1145/2484028.2484092`. 2.2

Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth J. F. Jones. An interface for agent supported conversational search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 452–456, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368926. doi: 10.1145/3343413.3377942. URL `https://doi.org/10.1145/3343413.3377942`. 1

Melanie Kellar, C. Watters, and M. Shepherd. A field study characterizing web-based information-seeking tasks. *J. Assoc. Inf. Sci. Technol.*, 58:999–1018, 2007. 2.4, 4.1.2

Tom Kenter and Maarten de Rijke. Attentive memory networks: Efficient machine reading for conversational search, 2017. 1

J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1):4–11, 1960. 1, 2.1

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Query reformulation using query history for passage retrieval in conversational search, 2020. 1, 2.2, 2.4

Cun Mu, Jun Zhao, Guang Yang, Binwei Yang, and Zheng Yan. Fast and exact nearest neighbor search in hamming space on full-text search engines, 2019. 2.3

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and characterizing user intent in information-seeking conversations. *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Jun 2018. doi: 10.1145/3209978.3210124. URL `http://dx.doi.org/10.1145/3209978.3210124`. 2.2

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. pages 117–126, 03 2017. doi: 10.1145/3020165.3020183. 1, 2.1, 5.1

Ning Sa and Xiaojun Jenny Yuan. Examining users' partial query modification patterns in voice search. *J. Assoc. Inf. Sci. Technol.*, 71(3):251–263, 2020. doi: 10.1002/asi.24238. URL `https://doi.org/10.1002/asi.24238`. 2.2

Mark Sanderson and W. Croft. The history of information retrieval research. *Proceedings of The IEEE - PIEEE*, 100:1444–1451, 05 2012. doi: 10.1109/JPROC.2012.2189916. 1

Research Department Statista. Alexa-compatible smart home devices 2017-2019, Feb 2020. URL `https://www.statista.com/statistics/912893/amazon-alexa-smart-home-compatible/`. 1

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. Question Rewriting for Conversational Question Answering. *arXiv e-prints*, art. arXiv:2004.14652, April 2020. 2.4

Sophie Walboomers and Claudia Hauff. A qualitative analysis on query reformulation types in conversational search scenarios, June 2020. 2.2

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. Few-shot generative conversational query rewriting, 2020. 1, 2.4