Predicting User Satisfaction with Intelligent Assistants

Julia Kiseleva[1,*] Kyle Williams[2,*] Ahmed Hassan Awadallah[3]
Aidan C. Crook[3] Imed Zitouni[3] Tasos Anastasakos[3]

[1]Eindhoven University of Technology, j.kiseleva@tue.nl
[2]Pennsylvania State University, kwilliams@psu.edu
[3]Microsoft, {hassanam, aidan.crook, izitouni, tasos.anastasakos}@microsoft.com

ABSTRACT

There is a rapid growth in the use of voice-controlled intelligent personal assistants on mobile devices, such as Microsoft's Cortana, Google Now, and Apple's Siri. They significantly change the way users interact with search systems, not only because of the voice control use and touch gestures, but also due to the dialogue-style nature of the interactions and their ability to preserve context across different queries. Predicting success and failure of such search dialogues is a new problem, and an important one for evaluating and further improving intelligent assistants. While clicks in web search have been extensively used to infer user satisfaction, their significance in search dialogues is lower due to the partial replacement of clicks with voice control, direct and voice answers, and touch gestures.

In this paper, we propose an automatic method to predict user satisfaction with intelligent assistants that exploits all the interaction signals, including voice commands and physical touch gestures on the device. First, we conduct an extensive user study to measure user satisfaction with intelligent assistants, and simultaneously record all user interactions. Second, we show that the dialogue style of interaction makes it necessary to evaluate the user experience at the overall task level as opposed to the query level. Third, we train a model to predict user satisfaction, and find that interaction signals that capture the user reading patterns have a high impact: when including all available interaction signals, we are able to improve the prediction accuracy of user satisfaction from 71% to 81% over a baseline that utilizes only click and query features.

Keywords: intelligent assistant, user satisfaction, user study, user experience, mobile search, spoken dialogue system

1. INTRODUCTION

Spoken dialogue systems have been thoroughly studied in the literature [37, 46–48]. However, it has only been in recent years that a new generation of intelligent assistants, powered by voice, such as Apple's Siri, Microsoft's Cortana, Google Now, have become common and popular on mobile devices. One of the reasons for the increased adoption is the recent significant improvement in accuracy of automatic speech recognition [38]. Intelligent assistants support multiple scenarios ranging from web search to proactive user recommendations [42]. In this work, we focus on dialogue mode of interaction with intelligent assistants. In this mode, a conversation takes place between the user and the intelligent assistant: the user speaks to the intelligent assistant, it responds and user speaks back, frequently referring to the subject of the previous request. This method of interaction is a more natural way for people to communicate and is often faster and more convenient (e.g., while driving) than typing. We call this type of interaction with intelligent assistants—*search dialogue*.

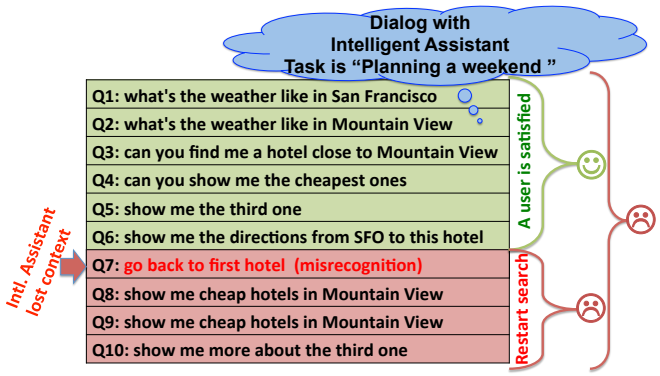

Figure 1: Example of search dialogue with intelligent assistant.

In search dialogue, users go through a sequence of steps in order to reach a desired goal: they solve one or more tasks, each of which consists of one or more search queries. As an example, consider the user dialogue in Figure 1: our user is trying to arrange a weekend in San Francisco. She has many tasks, from checking the weather to finding a hotel, or finding directions, etc. The user is engaged in a 'true' dialogue, i.e. the context is carried over across queries. When the intelligent assistant loses this context on Q_7, the user has to repeat some of the queries to rebuild the context and most probably gets dissatisfied with the intelligent assistant. So search dialogues are complex interactions, powered by voice control, with longer sessions consisting of different tasks and changes of focus within the same context. This is very different from traditional search in the query-response paradigm, and here session context becomes of crucial importance.

Clearly, evaluation of user satisfaction is an essential part of development of any intelligent assistant, as well as any traditional web search application. The ability to measure user satisfaction provides an understanding of the direction to take in order to improve the system. We can see from the example in Figure 1 that user satisfaction with search on intelligent assistants makes sense only for the entire dialogue, not as satisfaction with each query of a dialogue separately.

This prompts the need to better understand how users interact

SIGIR'16, July 17–21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911521

with search dialogue and how to define success and failure in terms of user experience—when are users (dis)satisfied? More specifically, we want to understand how we can measure and predict user satisfaction with search dialogue in ways that reflect perceived user satisfaction, and whether we can use traditional methods of offline and online evaluation or need to take other factors into consideration. The common practice for evaluating is to create a 'gold' standard (set of 'correct' answers) judged by editorial judges [22]. In the case of search dialogue, there may be no general 'correct' answer since the answers are highly personalized and contextualized (e.g. to a user's location or a user's past searches) to match better user-information needs. Another way to evaluate web search performance is through the use of implicit relevance feedback such as clicks, query length and landing page dwell time [2, 13, 17, 25, 26].

User satisfaction is widely adopted as a subjective measure of the quality of the search experience [27]. We know that user satisfaction for mobile web search is already very different when compared to desktop search [34]. The case of search dialogue is even more challenging for the measurement of user satisfaction [24]. Due to voice input-output to obtain answers directly from search dialogue without clicking, implicit relevance signals become far more important . The use of voice commands leads to a substantial increase in the length of queries: from 3.26 terms per query on average for mobile search to 4.48 for search dialogue, while also dramatically lowering the number of clicks per Search Engine Result Page (SERP): from 0.67 to 0.30[1]. Previous work [24] has modeled user satisfaction with intelligent assistants using generic explicit interaction signals (e.g. clicks, intelligent assistants request and response features, etc.) to simulate mobile search tasks, but the characteristics of more complex interactions and important touch-based signals were left unexplored. And [32] investigated self-reported user satisfaction in a related user study with a range of intelligent assistant tasks: device control, web search, and search dialogue. In this paper, we encompass all touch-based physical gestures that control the *mobile viewport location* (visible region on the mobile device), and screen taps (clicks), for the purpose of inferring user satisfaction with search dialogue. Concretely, our main research problem is:

> *How can we automatically predict user satisfaction with search dialogues on intelligent assistants using click, touch, and voice interactions?*

We breakdown our general research problem into three specific research questions.

> **RQ1:** *How can we define user satisfaction with search dialogues?*

As we show in Figure 1, a search dialogue is a sequence of user queries where each query is a step towards user satisfaction or frustration. We analyze interactions within search dialogue gradually increasing complexity of tasks and look at satisfaction with tasks.

> **RQ2:** *How can we predict user satisfaction with search dialogues using interaction signals?*

Clicks in web search have been extensively used to infer user satisfaction but clicks in search dialogue have lower significance due to the use of voice control and direct answers that does not require users to click. More insights can be gained by considering other interaction signals that characterize physical interaction with mobile devices. We investigate whether users' touch interactions provide useful signals for modeling user satisfaction for search dialogue and if they are more effective than using of general query, session, and click-based features.

> **RQ3:** *Which interaction signals have the highest impact on predicting user satisfaction with search dialogues?*

We analyze if touch-based features are important, while training an interaction-based predictor of satisfaction for search dialogue. Furthermore, we investigate which interaction signals are more important to predict user satisfaction by performing a correlation analysis between the interaction features. To answer our research questions, we set up a lab study with realistic tasks [5] for search dialogue derived from real user logs of a commercial intelligent assistant, measuring a wide range of aspects of user satisfaction. We use the outcome of the user study to understand and predict user satisfaction with intelligent assistants.

The remainder of this paper is organized as follows. Section 2 describes earlier work and background. We define user satisfaction though interaction signals for search dialogues in Section 3. Then, Section 4 introduces an approach for modeling user interaction with search dialogues. Section 5 provides a detailed description of the user study design to gather satisfaction labels. Finally, Section 6 reports our results, findings, and limitations. We conclude and discuss possible extensions of the current work in Section 7.

2. BACKGROUND AND RELATED WORK

This paper is relevant to three broad strands of research. First, we discuss research on spoken dialogue systems which are predecessors of the current intelligent assistants on mobile devices. Second, our work is related to evaluation of search quality because we propose a new model to evaluate user satisfaction with search dialogues. Third, our work is closely connected to the previous studies about user satisfaction in web search systems because we suggest a way to define and predict user satisfaction for search dialogues.

Spoken Dialogue Systems The main difference between traditional web search and intelligent assistants is their conversational nature of interaction. In the conversation mode of intelligent assistant, the technology can refer to the previous users' requests in order to understand the context of a conversation. For instance, in Figure 1 by asking Q_4 the user assumes that the intelligent assistant will 'know' that she is still interested in 'hotels in Mountain View'. Therefore, spoken dialogue systems [37] are closely related to intelligent assistants. Spoken dialogue systems understand and respond to the voice commands in a dialogue form; this area has been studied extensively over the past two decades [46–48]. Most of these studies focused on systems that have not been deployed in a large scale and hence did not have the necessary means to study how users interact with these systems in real-world scenarios, which led to most of the effort in evaluating spoken dialogue systems focusing on offline evaluation. Moreover, intelligent assistants on mobile devices support multiple scenarios of use compared with traditional spoken dialogue systems. For example, in addition to voice system response, intelligent assistants on mobile devices provide web search results, direct answers or proactive recommendations [42]. From these perspectives, intelligent assistants are similar to multi-modal conversational systems [20, 49].

This work is different from previous work on spoken dialogue systems in that, we study intelligent assistants on mobile devices and focus on analysis of user behavior that allows us to evaluate the system in an online setting, as well as identify instances of dissatisfaction with the system performance.

Search Quality Evaluation Historically, the key objective of information retrieval systems is to retrieve relevant information, typically in the form of documents or references to documents [40, 41].

[1]Statistics are calculated based on two weeks traffic of a commercial intelligent assistant in July 2015

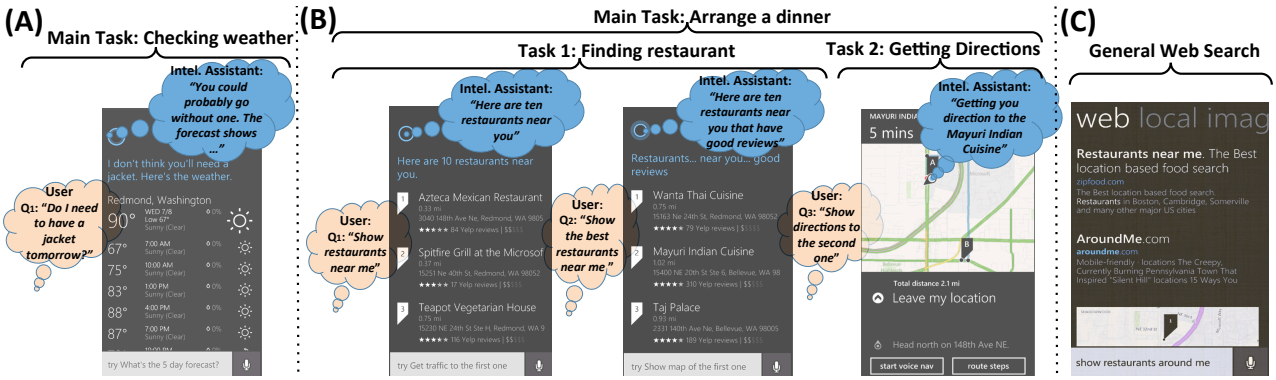

Figure 2: Examples (A) and (B) represent different types of intelligent assistant's response structured single task search dialogue and structured multi-task search dialogue accordingly. An example (C) represents a general SERP on mobile device.

In the simplest form, relevance can be defined as a score for a query-document pair. Given this query-document relevance score, many metrics have been defined, such as MAP, NDCG, DCG, MRR, P@n, TBG, etc. [22]. For such a setup, we have a collection of documents and queries that are annotated by human judges; such a setup is commonly used at TREC[2].

Recently online controlled experiments, such as A/B testing, have become widely used technique for controlling and improving search quality based on data-driven decisions [33]. This methodology has been adopted by many leading search companies such as Bing [7], Google [45], Facebook [4], and Yandex [10]. An A/B test is designed to compare two variants of a method (e.g. ranking on SERP, ads ranking , etc.) at the same time by exposing them to two user groups and by measuring the difference between them in terms of a *key metric* (e.g. the revenue, the number of visits, etc.), also known as an overall evaluation criterion. There are many existing studies towards better online evaluation which were devoted to inventing new metrics [9, 11] or improving existing ones [10]. The main goal of these studies was to make these metrics more consistent with the long-term goals [33]. User engagement metrics show different aspects of user experience. For instance, they can reflect **(1)** *user loyalty* – the number of sessions per user [43], **(2)** *user activity* – the number of visited web pages [35] or the absence time [11]. The periodicity engagement metrics of user behavior, which resulted from the Discrete Fourier transform of state-of-the-art engagement measures were applied in [9].

Our work is related to the online evaluation line of work since our objective is to build models that can be used to evaluate intelligent assistants, possibly in A/B testing settings. Our work is different in that we do not focus on how to run A/B experiments, rather we only focus on creating models that can be used to predict satisfaction.

User Satisfaction User satisfaction is widely adopted as a subjective measure of search experience. Kelly [27] proposes a definition: *'satisfaction can be understood as the fulfillment of a specified desire or goal'*. Furthermore, recently researchers studied different metrics reflective of user satisfaction, such as effort [52], and it has been shown that user satisfaction at the query-level can change over time [30, 31] due to some external influence. These changes lead to the necessity of updating the data collection. *Query-level* satisfaction metrics ignore the information about users' 'journey' from a question to an answer which might take more than one query [23]. Al-Maskari et al. [3] claim that *query-level* satisfaction is not applicable for informational queries. Users can run follow-up queries if they are unsatisfied with the returned results; reformulations can lead users to an answer – this scenario is called *task-level* user satisfaction [8, 17]. Moreover, Kelly et al. [29] have provided evidence that the most complex search tasks were similar to the work [6] characterization of complex tasks with respect to having multiple interdependent parts that needed to be addressed separately.

Previous research proposed different methods for identifying successful sessions. Hassan et al. [17] used a Markov model to predict success at the end of the task. Ageev et al. [1] exploited an expertise-dependent difference in search behavior by using a Conditional Random Fields model to predict a search success. Authors used a game-like strategy for collecting annotated data by asking participants to find answers to non-trivial questions using web search. On the other hand, situations when users are frustrated have also been studied. Feild et al. [12] proposed a method for understanding user frustration. Hassan et al. [18] and Hassan Awadallah et al. [19] have found that high similarity of successive queries is an indicator of an unsuccessful task. Our work is different from this line of work in that we focus on intelligent assistants while all these methods focus on analyzing user behavior when users interact with traditional search systems.

Most recently, user satisfaction for intelligent assistants on mobile devices started to gain attention [24]. Jiang et al. [24] focused on simulated tasks for device control and web search, and identify satisfactory and unsatisfactory sessions based on features used in predicting satisfaction on the web, as well as acoustic features of the spoken request. They do not focus on complex search dialogues and use use generic signals commonly used in Web search satisfaction modeling such as clicks and queries.

Sometimes, the information displayed on a SERP is sufficient to satisfy the users' information need. This phenomenon is called *good abandonment* [36, 44, 51] and was studied in [16] for mobile devices. The authors modeled viewing behavior based on touch interaction, and demonstrated the correlation of document relevance and *viewport* changing patterns on touch-enabled mobile devices. Recent research by Lagun et al. [34] extended this line of research to model the viewport for inferring user attention and satisfaction with SERPs. The absence of clicks is an emerging problem for intelligent assistants as well because they are frequently controlled by voice input.

Wildemuth et al. [50] reviewed over a hundred interactive information retrieval studies in terms of task complexity and difficulty, and found that the number of tasks and the number of facets were the main dimensions of task complexity. Recently, Kelly [28] linked perceived task complexity and effort, suggesting that user satisfaction may depend on the amount of effort to complete a complex task.

Our work focuses on modeling user satisfaction for intelligent assistants. We specifically focus on complex types of interaction

[2]Text REtrieval Conference: http://trec.nist.gov/

—search dialogue. We show that interaction signals are essential to infer user satisfaction with search dialogue and demonstrate how they can be used in practice. We also focus on studying new interaction signals (such as touch and viewport changes) to model user's attention. We introduce a general notion of user satisfaction and exploit an extended list of interaction signals in order to predict user satisfaction with search dialogue

To summarize, the key distinctions of our work compared to previous efforts are: we studied a new method of user interaction with intelligent assistants on mobile devices, search dialogue, and we proposed a method to measure and predict user satisfaction for search dialogue using touch-interaction signals. Our metric is applicable to evaluation both online (e.g., introducing a new ranker or answer type for the intelligent assistant) and offline (e.g., mining search dialogues where users are dissatisfied).

3. DEFINING USER SATISFACTION

In this section we investigate **RQ1:** *How can we define user satisfaction with search dialogues?* In the case of search dialogue, the key distinction of this scenario is the ability of the intelligent assistant to maintain the context of the conversation. Moreover, responses provided by intelligent assistants can be either in the form of a structured answer or in the form of the usual mobile SERP. Figure 2 (A) and (B) illustrates examples of structured answers from a commercial intelligent assistant. Examples of tasks when this type of interaction is activated include requests about restaurants, hotels, travel, weather, etc. Structured answers differ significantly from the usual mobile SERP (e.g., Figure 2 (C)). We characterize different types of search dialogues based on our broad analysis of the logs from a commercial intelligent assistant (Section 3.1). We also present a generalized definition of user satisfaction with search dialogue using interaction signals (Section 3.2).

3.1 Search Dialogue Types

After intensive analysis of the logs of a commercial intelligent assistant, we split search dialogues into two types: single task search dialogues and multi-task search dialogues. Roughly 50-55% of interactions can be characterized as single task search dialogues, the rest as multi-task search dialogues.

Single Task Search Dialogue Single task search dialogue has one underlying atomic information need and mostly consists of one query and one answer. An example of a single task search dialogue is the weather-related information need shown in Figure 2 (A). Single task search dialogues are very similar to mobile web search and follow the query-response paradigm. We expect that they can be evaluated using *query-level* satisfaction.

Multi-Task Search Dialogue Multi-task search dialogue consists of multiple interactions with the intelligent assistant that lead towards one final goal e.g. 'plan a night out'. These long and complex interactions can be divided into a series of tasks. Obviously, multi-task search dialogues are more complex than other search dialogues because of a greater number of interactions whereby the user speaks to the intelligent assistant, the intelligent assistant responds, the user speaks back to it and so on.

An example of multi-task search dialogue is presented in Figure 2 (B): the intelligent assistant is used to arrange a dinner. The user makes the following transitions in this search dialogue:

- Q_1 : asking for a list of the nearest restaurants.
- Q_2 : sorting the returned list to find the best restaurants (***During the transition*** $Q_1 \rightarrow Q_2$, *the intelligent assistant 'knows' that the user is referring to the list of restaurants from the previous query).*

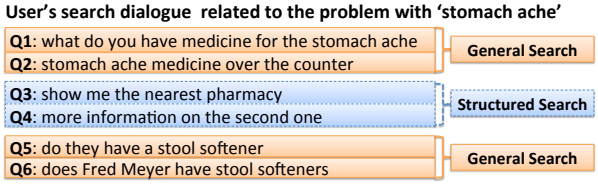

Figure 3: An example of the search dialogue where structured answer and general web SERP are used.

- Q_3 : selecting a restaurant from the list and asking for the directions (***During the transition*** $Q_2 \rightarrow Q_3$ ***:*** *the intelligent assistant 'knows' that the user is working with the sorted list of restaurants).*

We notice that some user needs turn out to be too complex to answer with the structured interface. An example where a user needs help with a stomach ache that is shown in Figure 3. In this case, the intelligent assistant used both general web search and structured dialogue interface to respond to the user's requests. The intelligent assistant redirects a user to general search if the intelligent assistant deems that general SERP will satisfy the user's information needs better such as for queries: Q_1, Q_2, Q_5 and Q_6 in Figure 3.

A search dialogue is not just a sequence of $\langle Q, SERP \rangle$ pairs consisting of the SERP returned by intelligent assistant in response to the voice query Q. Search dialogue consists of one or more tasks, each of which consists of one or more queries. To better understand requirements for the user study setup, we divide search dialogues into single- and multi-task. Our hypothesis is that it is important for evaluation of user satisfaction with intelligent assistants if a response to a voice query Q can be either in a structured form ($SERP^{str}$, see Figure 2 (A) and (B)) or in a form of a general web search ($SERP^{web}$, see Figure 2 (C)).

3.2 User Satisfaction with Search Dialogues

Based on our analysis of a commercial intelligent assistant logs we hypothesize that much of the frustration happens when the intelligent assistant is not able to maintain the context and users need to start their search over in order to complete their tasks. As we present in the example in Figure 1, the intelligent assistant lost the context in the transition $Q_6 \rightarrow Q_7$ due to an automatic speech recognition error, and the user had to start over. Overall user satisfaction with the search dialogue decreases dramatically in this case despite the fact that the user seemed to be satisfied with the previous transitions: $Q_1 \rightarrow \cdots \rightarrow Q_6$. Furthermore, it is likely to be especially frustrating since the mistake happens at the end of the session.

Single task search dialogue has one main task T that can be represented as follows: $T = \big(\langle Q_1, SERP_1 \rangle, \ldots, \langle Q_n, SERP_n \rangle\big)$. For any given task T, there are a set of interaction signals (e.g. touch, viewport change, etc.) that we denote as $I(T)$ and it can be defined as function f that combines all interactions for every $\langle Q, SERP \rangle$ pair in T:

$$I(T) = f\Big(I\big(\langle Q_1, SERP_1 \rangle\big), \ldots, I\big(\langle Q_n, SERP_n \rangle\big)\Big). \quad (1)$$

In the case of multi-task search dialogue, the search dialogue has more than one task and can be viewed as a sequence of tasks: $T_1, \ldots, T_m$. Interaction signals within the search dialogue are defined through the function g that aggregates user interaction over tasks happening during search dialogue:

$$I(T_1, \ldots, T_m) = g\big(I(T_1), \ldots, I(T_m)\big). \quad (2)$$

Our objective is to define a function h that given a set of interaction signals would predict whether the user was satisfied or not.

Table 1: Description of implicit features per search dialogue

Feature Name	Feature Description
F_1 *NumQueries*	Number of queries
F_2 *NumClicks*	Number of clicks
F_3 *NumSATClicks*	Number of clicks ($>$ 30 sec. dwell time)
F_4 *NumDSATClicks*	Number of clicks ($\leq$ 15 sec. dwell time)
F_5 *TimeToFirstClick*	Time (seconds) until the first click
F_6 *MetaphoneLevenstein*	Levenstein similarity between pronunciation and writing
F_7 *MetaphoneSubstring*	Substring similarity between pronunciation and writing

For multi-task search dialogues, h can be defined as:

$$SAT(T_1, \ldots, T_m) = h\big(I(T_1, \ldots, T_m)\big). \quad (3)$$

In the case of a single task search dialogue that consists of one task T, Equation 3 would be simplified to $SAT(T) = h\big(I(T)\big)$. If single task search dialogue consists of a single query, the equation can be further simplified to $SAT(T) = h\big(I(\langle Q, SERP\rangle)\big)$, like in standard query-level satisfaction.

In this section, based on extensive analysis of the logs of an intelligent assistant, we characterized search dialogues as single- and multi-task, divided queries as giving either a structured or a general web search response, and conceptually modeled user satisfaction with search dialogues. Additionally we illustrated that the overall user satisfaction with search dialogue cannot be reduced to the query or even task level satisfaction, because of the dependency between them and the expectation that the intelligent assistant maintains the context during the whole interaction within a dialogue.

4. MODELING USER INTERACTIONS

This section addresses **RQ2:** *How can we predict user satisfaction with search dialogues using interaction signals?* First, we describe used interaction signals that are logged as the following two types of features: (1) general *implicit features* which have been used in previous work on characterizing user behavior with general Web search [2, 13, 17, 25] and intelligent assistants [24] (Section 4.1), and (2) *touch and attention features* which, we believe, provide a different perspective for modeling satisfaction with search dialogues (Section 4.2). Note that some of these features were also shown to be useful for predicting the relevance of web search results [15, 16, 34]. These two types of features are used to define $I(\langle Q, SERP\rangle)$ which is a component of Equation 1. Finally, we present a method for modelling user interaction with the search dialogue task T to represent $I(T)$ from Equation 1 (Section 4.3).

4.1 Implicit Features

Table 1 lists the utilized implicit features: $(F_1, \ldots, F_7)$.

Queries and Click Features $(F_1, \ldots, F_5)$**:** In our case *click* means tapping a result item (e.g., the best answer from a list of candidates). We use the following features that are calculated across the entire search dialogue task: the number of queries (F_1) the number of clicks (F_2), the number of satisfied clicks, defined as clicks with dwell time $>$ 30 seconds (F_3), as well as the number of dissatisfied clicks, defined as clicks with dwell time $\leq$ 15 seconds (F_4), and the total time (seconds) before the first click in search dialogue (F_5). Note that previous work [13] has shown that long dwell time clicks ($>$ 30 seconds) are highly likely to indicate satisfaction while quick-back clicks ($\leq$ 15 seconds) are highly likely to indicate dissatisfaction.

Table 2: Description of touch features per search dialogue

Feature Name	Feature Description
F_9 *NumSwipes*	Number of Swipes
F_{10} *NumUpSwipes*	Number of up-swipes
F_{11} *NumDownSwipes*	Number of down-swipes
F_{12} *SwipedDistance*	Total distance swiped (pixels)
F_{13} *AvgNumSwipes*	Number of swipes normalized by time
F_{14} *AvgSwipeDistance*	Total distance divided by num. of swipes
F_{15} *DistanceByTime*	Total swiped distance divided by time
F_{16} *DirectionChanges*	Number of swipe direction changes
F_{17} *DurationPerAns*	SERP answer duration (seconds) which is shown on screen (even partially)
F_{18} *FractionPerAns*	Fraction of visible pixels belonging to SERP answer
F_{19} *ReadTimePerAns*	Attributed time (seconds) to viewing a particular element (answer) on SERP
F_{20} *1DReadTimePerPix*	Attributed time (seconds) per unit height (pixels) associated with a particular element on SERP
F_{21} *2DReadTimePerPix*	Attributed time (milliseconds) per unit area (square pixels) associated with a particular element on SERP

Acoustic Features (F_6, F_7)**:** We utilize acoustic feature to characterize voice interaction happening in search dialogues. More specifically, we use the phonetic similarity between consecutive requests to identify patterns of repetition. Metaphone representation [39] is a way of indexing words by their pronunciation that allows us to represent words by how they are pronounced as opposed to how they are written. Phonetic similarity is assessed by computing the edit distance between the Metaphone representation of two utterances. For example, a voice query *'WhatsApp'* may be incorrectly recognized as *'what's up'*, but their metaphone codes are both *'WTSP'*. In such cases, this phonetic similarity feature helps us detect repeated or similar requests that are missed by normal text similarity features based on recognized speech. As similarities metrics we use Levenstein Distance (F_7) and Substring (F_8).

4.2 Touch Features

One of the main contributions of this work is the introduction of touch and attention features for detecting user satisfaction with search dialogues. We focus on touch-based features related to the way in which users interact with the screen and features based on elements visible to users. This serves as a surrogate for what the user is paying attention to on the page and how this changes throughout the search dialogue. Table 2 lists the used touch features. Capturing touch events is not easy in practice because of non-standard instrumentation [21]. We derive interaction features and the exact information that was displayed on the phone screen at any given time using *mobile viewport logging*. This allows us to record the portion of the answer/result currently visible on the screen, as well as bounding boxes of all results shown on the page. For instance, if an element is visible in the viewport at some point in time and then no longer visible, one can infer that a gesture must have taken place. Furthermore, if an element below the original element becomes visible, then one can infer that it must have been a downward swipe action. We use element-tracking in the viewport to infer features related to swipes happening during search dialogue: $F_9, \ldots, F_{16}$.

Lagun et al. [34] showed that there is strong correlation between the time for which a result is visible and its gaze time. Following this observation, we approximate how much attention different SERP elements get. Features F_{17}, F_{18} are used to characterize vis-

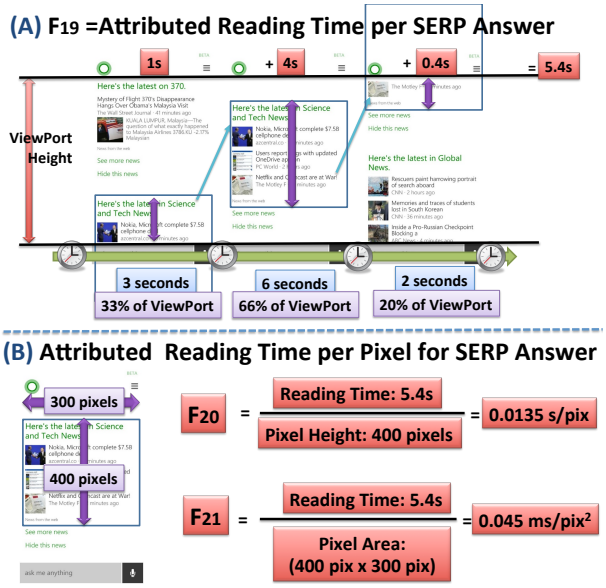

Figure 4: The illustration how to capture (A): F_{19} *ReadingTimePerAnswer* and (B): F_{20}, F_{21} *ReadingTimePerPixel.*

ibility of SERP answers. The feature F_{19} attempts to attribute the time the user spends looking at each stationary viewport to the different elements based on their area. Features F_{20} and F_{21} are responsible for reading time per pixel, they normalize the attributed reading time so that size of the content region does not introduce a systematic weight into the metric. Figure 4 illustrates how F_{19} is captured in the example (A) and how F_{20} and F_{21} are calculated in the example (B). To aggregate the features $F_{17}, \ldots, F_{21}$ at the $\langle Q, SERP \rangle$-level, we use four types of aggregation: average (Avg), maximum (Max), minimum (Min), and standard deviation ($Stdev$).

We presented a list of implicit and touch features that are collected on $\langle Q, SERP \rangle$-level during user interaction with search dialogues on intelligent assistants. We define $I(\langle Q, SERP \rangle)$ by the feature vector: $(F_1, \ldots, F_{21})$. Next, we will explain how to model user interaction on the task level—$I(T)$.

4.3 User Interactions over Search Dialogue

We showed that search dialogue tasks have underlying semantic structure and potentially can be divided into single task search dialogues and multi-task search dialogues. There is no automatic search dialogue analyser available so we cannot split search dialogues into tasks on the fly. The goal of this work is to deliver online metric for user satisfaction with search dialogues on intelligent assistants. Potentially, the semantic structure of a search dialogue task is not entirely flat and it might have a tree structure. Developing an automatic tool to mine the search dialogue structure is a promising direction for future work.

The intelligent assistant has two types of responses to a voice query Q: either in a structured form, $SERP^{str}$ as illustrated in Figure 2 (A) and (B), or in the form of a general web search, $SERP^{web}$ as illustrated in Figure 2 (C). Our hypothesis is that the type of response of intelligent assistants can be used to approximately divide search dialogues into the different types of tasks. Our assumption relies on the internal logic of the intelligent assistant that returns $SERP^{str}$ when tasks are about different types of locations (restaurants, hotels, pharmacies, shops etc.), directions to locations, or weather. If the intelligent assistant deems that information from general web is more suitable for a query then it returns $SERP^{str}$. This kind of intelligent assistant response still differs from general mobile search because it looks like a dialogue. For example, if a user voice query can be answered using the knowledge graph then, the intelligent assistant speaks the answer out aloud.

We define the function f from Equation 1 through aggregation. Further, in our experiment we use geometric mean as aggregation. We experimented with other aggregation functions, and they yielded similar or worse performance. We apply three techniques to define $I(T)$ for the search dialogue consisting of n queries in total, m queries resulted in $SERP^{str}$ and k queries resulted in $SERP^{web}$:

A_1: considering only interaction with $\langle Q, SERP^{str} \rangle$:

$$I(T) = \Big(\prod_{i=1}^{m} I\big(\langle Q_i, SERP_i^{str} \rangle\big) \Big)^{1/m}; \tag{4}$$

A_2: considering interactions with all $\langle Q, SERP \rangle$ equally:

$$I(T) = \Big(\prod_{i=1}^{n} I\big(\langle Q_i, SERP_i \rangle\big) \Big)^{1/n}; \tag{5}$$

A_3: separating interactions with $\langle Q, SERP^{str} \rangle$ and $\langle Q, SERP^{web} \rangle$ as two different tasks T^{str} and T^{web}:

$$I(T^{str}, T^{web}) = \Big[\Big(\prod_{i=1}^{m} I\big(\langle Q_i, SERP_i^{str} \rangle\big) \Big)^{1/m}, \Big(\prod_{j=1}^{k} I\big(\langle Q_j, SERP_j^{web} \rangle\big) \Big)^{1/k} \Big]. \tag{6}$$

In this section we introduced the list of features to model user interactions. We focused specifically on presenting interaction signals which are promising for modeling user interaction with intelligent assistants. Next, we will describe the set up for our lab study with real-world tasks for search dialogue derived from real user logs. The outcome of the study will be used to understand how important interaction signals are for modeling user satisfaction with search dialogue.

5. USER STUDY

This section describes the design of the user study to collect user interactions for search dialogues. The collected data is used to investigate our research questions: **RQ2–3**. While designing tasks for our user study, we rely on the following requirements: (1) the suggested tasks should be realistic; (2) following Borlund [5] we construct the tasks so that participants could relate to them and they would provide '*enough imaginative context*.'

Participants We recruited 60 participants to participate in the study. All participants were college or graduate students residing in the United States. They all had good command of English. 75% of participants were male and the remaining 25% were female. The average age of participants is 25.5 ($\pm$ 5.4) years. They were reimbursed $10 gift card for participating in the study.

Tasks We analyzed over 400,000 search dialogues from the search logs of a commercial intelligent assistant to generate tasks for user study. Based on our analysis we generated **eight** tasks for the user study that were designed to cover approximately 70-80% of subjects queried by real users of the intelligent assistant. We formulate tasks in a free form in order to encourage query diversity and stimulate either genuine satisfaction or frustration with returned results. The final tasks for user study consist of:

- **one** single task search dialogue that is related to the weather e.g. **Task A:** *'Check if you need to wear a coat tomorrow?'*;
- **four** multi-task search dialogues that include two subjects e.g. **Task B:** *'You are planning night out. Pick a restaurant based on your preferences: cheap, best review, or closest. Find out driving directions to the selected restaurant.*;
- **three** multi-task search dialogues that require at least three switches within the same context e.g. **Task C:** *'You are planning vacation. Pick a place. Check if the weather is good enough for the period you are planning the vacation. Find a hotel that suits you. Find out driving directions to this place.*

For each task, we recorded an audio that verbally described the task objective. Following the study [24], we did not show the participants the written description while they were working on the task as it was demonstrated many participants directly used the sentences shown in descriptions as requests. We strongly wanted to avoid such outcome because our goal was simulate real user behavior.

Study Setup Participants performed the tasks on a mobile phone with a commercial intelligent assistant installed. If the task needed access to specific device resources, functions or applications (e.g. maps), they were pre-installed to make sure users would not encounter problems. The experiment was conducted in a quiet room, so as to reduce the disturbance of external noise. Although the real environment often involves noise and interruption, we eliminated those factors to simplify the experiment. While participants were doing the user study all their interactions were logged using an internal API.

The participants watched a 4 minutes video with instructions[3] that explained how to use the intelligent assistant. Then, participants worked on **one** training task and **eight** formal tasks. We instructed participants that they should stop a task when they had accomplished their goal or if they became frustrated and wanted to give up. After completing each task, participants were asked to answer the following four questions:

1. *Were you able to complete the task?*
2. *How satisfied are you with your experience in this task in general?*

   In case of multi-task search dialogue participants indicate their graded satisfaction e.g. for **Task C** questions were:

   2.1 *How satisfied are you with finding a hotel?*
   2.2 *How satisfied are you with finding a direction?*
3. *Did you put in a lot of effort to complete the task?*
4. *How well did the intelligent assistant recognize your voice?*

Except for the first question which required a Yes/No answer, all questions were answered using a 5 point likert scale. Additionally, to stimulate participants' involvement in search dialogues, we asked them to answer clarifying question(s) about task output. For example, if the task was about finding a restaurant the participant would need to indicate its name in the questionnaire. The total experiment time was about 30 minutes.

User Study Summary We stimulated participants' involvement by giving free form tasks. They needed to formulate their own goals for the task and it leads to satisfaction or frustration. For example, out of 60 responses for the **Task C** we extracted 46 references to unique places. As a result of free task formulation we obtained a diverse query set, characterized as follows: in total, participants perform 540 tasks that involved $2,040$ queries in total of which $1,969$ were unique; the average query length was 7.07. The single task search dialogue as **Task A** generated 130 queries in total, four multi-task search dialogues as **Task B** generated 685 queries, and three multi-task search dialogues as **Task C** generated $1,355$ queries.

[3]Link to instructions withheld to preserve anonymity.

6. RESULTS AND FINDINGS

We now investigate our **RQ3:** *Which interaction signals have the highest impact on predicting user satisfaction with search dialogues?* We begin by introducing our results on the prediction quality of user satisfaction with search dialogues (Section 6.1). We conclude by presenting a correlation analysis between the interaction features and user satisfaction (Section 6.2).

6.1 Predicting User Satisfaction

The purpose of this study is to predict overall user satisfaction with search dialogues. Therefore we do not utilize graded satisfaction in this work but it would be useful for future research. In our user study, users reported overall satisfaction using a 5 point likert scale. Due to the large difference in rating distributions between the single- and multi-task search dialogue we consider the evaluation as a binary classification problem. We divide the labeled search dialogues into binary classes: satisfied (SAT) – users provided 5 or 4; dissatisfied (DSAT) – everything else. This resulted in the following proportion of positively and negatively labeled search dialogues: SAT – 64% and DSAT – 36%.

We formulate a supervised classification problem where, given a search dialogue, the goal is to classify it to SAT or DSAT. We train Gradient Boosted Decision Trees (GBDT) [14] as a satisfaction predictor h presented in Equation 3. We experiment with other classifiers (logistic regression, SVM), and they yield similar or worse performance. Hence we only report the results of GBDT. We use 10-fold cross validation. For each training fold, we use grid search to optimize for the number of leaves, tree depth, and number of leaves required to split. We train our predictors based on different subsets of features from $(F_1, \ldots, F_{21})$. For each experiment we report the overall accuracy (Acc), average F_1 score (Avg. F_1), area under the curve (AUC); and precision (P), recall (R) and F_1 score (F_1) for SAT and DSAT separately. The results are shown in Table 3.

The baseline is the classifier trained on queries and click features which are aggregated over a search dialogue using Equation 4. We observe that the baseline is overly optimistic with a low DSAT recall (30%) and high SAT recall (93%), showing that it is effective in picking up the imbalance in SAT/DSAT distribution but far less effective in distinguishing satisfaction from dissatisfaction. We train the predictor P_1 on an expanded feature set, adding the Methaphone features (F_6, F_7). From Table 3, we can see that the predictor P_1 shows statistically significant improvement ($p < 0.05$) in Acc , SAT P, DSAT R, DSAT F_1, Avg. F_1 and AUC when compared against the baseline. Next, we expand feature set by adding the touch signals from Table 2.

We use the three proposed techniques for feature aggregation over task(s) while training based on $(F_1, \ldots, F_{21})$: A_1 (Equation 4) for the predictor P_3, A_2 (Equation 5) for P_4, and A_3 (Equation 6) for P_5. Based on results in Table 3, we can infer that the predictors P_2, P_3 and P_4 demonstrate statistically significant improvements ($p < 0.05$) in Acc, SAT P, DSAT R, DSAT F_1, Avg. F_1 and AUC when compared against the baseline, indicating that the touch features incorporated in prediction models are fundamental to evaluation of user satisfaction with search dialogues. Also from Table 3, we can infer that the aggregation A_3 (when we separate user interactions: $SERP^{str}$ and $SERP^{web}$) is most beneficial

Table 3: Measurements of prediction quality based on different subsets of features. The relative improvements compared to the baseline are provided in parentheses. * indicates statistical significant ($p < 0.05$) using paired t-tests compared to the baseline.

Features Description	Accuracy (%)	SAT (%)			DSAT (%)			Avg. F_1 (%)	AUC (%)
		Precision	Recall	F_1	Precision	Recall	F_1		
Baseline: $A_1(F_1, \ldots, F_5)$ (Eq. 4)	70.62	70.72	**92.91**	80.31	70.50	30.37	42.45	61.38	61.51
$\mathbf{P}_1$: $A_1(F_1, \ldots, F_7)$ (Eq. 4)	78.53* (+11.20)	81.81* (+15.68)	85.73 (-7.73)	83.72 (+4.25)	71.24 (+1.76)	65.55* (+115.84)	68.51* (+61.37)	76.11* (+24.00)	81.20* (+32.01)
$\mathbf{P}_2$: $A_1(F_1, \ldots, F_{21})$ (Eq. 4)	78.78* (+11.55)	80.98* (+14.51)	87.75 (-5.55)	84.23 (+4.88)	74.69 (+5.94)	62.61* (+106.16)	68.12* (+60.46)	76.17* (+24.10)	83.59* (+35.90)
$\mathbf{P}_3$: $A_2(F_1, \ldots, F_{21})$ (Eq. 5)	80.21* (+13.58)	82.55* (+16.73)	87.99 (-5.30)	**85.18 (+6.07)**	**76.28*** **(+8.20)**	66.07* (+117.55)	70.81* (+66.80)	78.00* (+27.07)	83.31* (+35.44)
$\mathbf{P}_4$: $A_3(F_1, \ldots, F_{21})$ (Eq. 6)	**80.81*** **(+14.43)**	**84.89*** **(+20.04)**	85.42 (-8.06)	85.15 (+6.03)	73.45 (+4.18)	**72.55*** **(+138.89)**	**73.00*** **(+71.95)**	**79.08*** **(+28.83)**	**85.62*** **(+39.20)**

Table 4: Pearson correlations between satisfaction (SAT) and implicit features. Results are statistically significant ($p < 0.05$)

Feature Type	Correlation
$F_7(Q_i, Q_{i+1})$ *[MetaphoneSubstring]*	0.45
F_4 *[NumDSATClicks]*	0.31
F_5 *[TimeToFirstClick]*	0.30
F_2 *[NumClicks]*	0.27
F_6 *[MetaphoneLevenstein]*	0.23
F_3 *[NumSATClicks]*	0.12
$F_7(Q_{i+1}, Q_i)$ *[MetaphoneSubstring]*	-0.16
F_1 *[NumQueries]*	-0.49

one when compared against the baseline. In the next subsection, we present features analysis to characterize the relative importance of different features.

6.2 Features Analysis

To understand the impact of implicit features $(F_1, \ldots, F_7)$ from Table 1, we calculate the Pearson correlation between the user satisfaction label (SAT) and each feature. The results are presented in Table 4. Feature $F_7(Q_i, Q_{i+1})$, which indicates that the subsequent query Q_{i+1} in the task contains prior query Q_i, is positively correlated with SAT. Expanding the query, or rather refining the query to better specify the intent, is a common user behavior and is expected to increase the probability of finding satisfactory content on the subsequent SERP. The complementary feature, $F_7(Q_{i+1}, Q_i)$, however, reflects the case where the subsequent query Q_{i+1} in the task is contained within the prior query Q_i; this feature is negatively correlated with SAT. Speech recognition errors in $Q_i \rightarrow Q_{i+1}$ can give rise to this type of feature, and the negative correlation is expected from such transitions. Our findings are similar to the previously reported results [24]. Based on relatively high correlation between click-based features $(F_2, \ldots, F_5)$ we infer that clicks during search dialogues can be interpreted as a sign of user satisfaction. We find that the search dialogue length, in terms of F_1, is negatively correlated with satisfaction. Long conversations can be result from two types of behaviors: (a) multiple attempts of users to have their speech properly recognized, or (b) the loss of context by the intelligent assistant during the conversation, forcing users to restart the conversation; both of these explain the observed negative correlation.

Table 5 shows the results of correlation analysis for the touch features $(F_8, \ldots, F_{21})$ using aggregations A_2 and A_3. We present the top 5 positively correlated features and the top 5 negatively correlated features. To explain the correlations, we present three hypotheses:

Table 5: Pearson correlations between satisfaction (SAT) and touch features. Results are statistically significant ($p < 0.05$)

Feature Type	Cor.
A_2 **(Eq. 5) for aggregating Touch Features**	
$Stdev(F_{18})$ *[FractionPerAns]*	0.23
$Min(F_{20})$ *[1DReadTimePerPix]*	0.20
$Stdev(F_{19})$ *[ReadTimePerAns]*	0.19
$Avg(F_{20}))$ *[1DReadTimePerPix]*	0.19
$Max(F_{20}))$ *[1DReadTimePerPix]*	0.18
...	...
F_{10} *[NumUpSwipes]*	-0.10
F_9 *[NumSwipes]*	-0.12
F_{11} *[NumDownSwipes]*	-0.12
F_{12} *[SwipedDistance]*	-0.13
F_{15} *[DistanceByTime]*	-0.18
A_3 **(Eq. 6) for aggregating Touch Features Aggregation**	
$I(\langle Q, SERP^{str}\rangle)$: $Max(F_{18})$ *[FractionPerAns]*	0.35
$I(\langle Q, SERP^{str}\rangle)$: $Stdev(F_{18})$ *[FractionPerAns]*	0.34
$I(\langle Q, SERP^{str}\rangle)$: $Max(F_{19})$ *[ReadTimePerAns]*	0.32
$I(\langle Q, SERP^{str}\rangle)$: $Avg(F_{18})$ *[FractionPerAns]*	0.31
$I(\langle Q, SERP^{str}\rangle)$: $Avg(F_{19})$ *[ReadTimePerAns]*	0.31
...	...
$I(\langle Q, SERP^{web}\rangle)$: $Min(F_{20})$ *[1DReadTimePerPix]*	-0.35
$I(\langle Q, SERP^{web}\rangle)$: $Stdev(F_{18})$ *[FractionPerAns]*	-0.28
$I(\langle Q, SERP^{web}\rangle)$: $Min(F_{18})$ *[FractionPerAns]*	-0.32
$I(\langle Q, SERP^{web}\rangle)$: $Avg(F_{18})$ *[FractionPerAns]*	-0.35
$I(\langle Q, SERP^{web}\rangle)$:$Max(F_{18})$ *[FractionPerAns]*	-0.35

$\mathbf{H}_1$: The SERP for a query is ordered by a measure of relevance as determined by the system, then additional exploration is unlikely to achieve user satisfaction, but is more likely an indication that the best-provided results (i.e. the SERP top) are insufficient to address the user intent.

$\mathbf{H}_2$: In the converse case of $\mathbf{H}_1$, when users find content that satisfies their intent, their likelihood of scrolling is reduced, and they dwell for an extended period on the top viewport.

$\mathbf{H}_3$: When users are involved in a complex task, they are dissatisfied when redirected to a general mobile SERP, as opposed to receiving an explicit structured answer from the intelligent assistant (e.g. the transition $Q_4 \rightarrow Q_5$ in Figure 3). Unlike $\mathbf{H}_2$, the absence of scrolling on this landing page is an indication of dissatisfaction.

The features in Table 5 are explained in more depth below. A

large $Stdev(F_{18})$ characterizes the situation where roughly half of the available answers are observed and the other half are not. This would occur when there is minimal or no scrolling behavior, since answers at the top of the SERP are visible and the answers toward the bottom are hidden from view. F_{20} is well-defined only for observable content, and when users do not scroll, this value will be identical for all items on the SERP. As such, in the absence of scrolling, $Min(F_{20})$ will be large, and therefore a positive correlation with SAT is consistent with our hypotheses. F_{19}, on the other hand, is well-defined for all answers, observed or not, but will be equal to zero for answers that are not observed. When there is minimal scrolling and a long dwell on the top viewport, F_{19} will be positive and large for the observed answers, and zero for the unobserved content, giving rise to a large $Stdev(F_{19})$. $Avg(F_{20})$ characterizes the same behaviour as $Min(F_{20})$ when users do not scroll at all, but, when users do scroll small distances, $Min(F_{20})$ would drop substantially whereas $Avg(F_{20})$ would remain relatively stable; a positive correlation with SAT is consistent with $\mathbf{H}_2$. A large $Max(F_{20})$ implies that users paused and dwelled on one portion of the page for an extended period, also consistent with $\mathbf{H}_2$.

Table 5 (A_2) shows that SAT is negatively correlated with (F_9, $\ldots$, F_{12}), which describe user swipes. Swipe down, up, or both are signs of exploration of the result set and a negative correlation of number of swipes and swipe-distance with SAT is consistent with $\mathbf{H}_1$. F_{15} provides a measure of the speed of exploration of the content. The observed negative correlation implies that fast swiping indicates dissatisfaction, and it is consistent with users who are skimming through and exploring the results without success, supporting $\mathbf{H}_1$. These results are consistent with the findings of [34], who concluded that scrolling is negatively correlated with SAT.

For the aggregation A_3 (Equation 6), we separate interaction with structured answers, $I\big(\langle Q, SERP^{str}\rangle\big)$, and interaction with general mobile SERP, $I\big(\langle Q, SERP^{web}\rangle\big)$. The correlation between SAT and F_{18}, F_{19} calculated though interaction with $SERP^{str}$ is even stronger. The same set of features calculated for interactions with $SERP^{web}$ is negatively correlated with SAT, which is consistent with $\mathbf{H}_3$. Users who are redirected to $SERP^{web}$ and does not scroll likely land there unintentionally, as a consequence of a voice-misrecognition or loss of context by the intelligent assistant. While Table 5 only shows the top features, the entire list of correlations for A_3 are consistent with the $\mathbf{H}_1$, in agreement with our previous finding for the aggregation A_2. Furthermore, we can see that swiping actions during interactions with $SERP^{web}$ have higher negative correlation than with $SERP^{str}$. Here, users are plausibly frustrated and perform quick swipes through $SERP^{web}$. The above observations lead us to the following conclusion—that users expect to find answers on the SERP without any 'additional effort' (e.g. scrolling), and users are not satisfied if the intelligent assistant cannot answer their request explicitly and redirects them to a general mobile SERP.

Although our paper shows that our method has a strong potential, there are at least two limitations that can be improved in future work. The first limitation is the collected data during user study which can be improved in terms of size and diversity. One way to do that is to monitor users as they do their normal tasks via additional instrumentation installed on their phones and prompt them to answer questions about their satisfaction. Another area of improvement is using data collected from multiple intelligent assistants. Most available intelligent assistants support search dialogues and the features we use are independent of the task subject and hence should be useful regardless of which tasks are supported by which assistants. Nevertheless, training and testing our models on data from different assistants can be very useful for proving their generality. This is particulate challenging though given the difficulty of performing third-party instrumentation on mobile devices.

To summarize, extensively experimenting with the user study data, we concluded that touch and attention based features are extremely helpful for predicting user satisfaction with intelligent assistants. Finally, we conducted feature analysis and concluded that active user interactions with the mobile device (e.g., scrolling) is the strong signal of user dissatisfaction with intelligent assistant.

7. CONCLUSIONS AND FUTURE WORK

The paper extends earlier work on desktop and general mobile search [16, 24, 34] and presents the first quantitative study for user satisfaction with the modern generation of intelligent assistants. Intelligent assistants allow for radically new means of information access: making a real dialogue with a context using voice commands and touch interactions. Evaluation of user satisfaction is crucial for intelligent assistants development. As the popularity of intelligent assistants rapidly grows, a strong need for better understanding and precise evaluating of user satisfaction grows correspondingly.

Our main research question was: *How can we automatically predict user satisfaction with search dialogues on intelligent assistants using click, touch, and voice interactions?* First, we studied **RQ1:** *How can we define user satisfaction with search dialogues?* We studied search dialogues by analyzing real logs of a commercial intelligent assistant and introduced two types of the dialogues: single task search dialogues and multi-task search dialogues. We also illustrated that the dialogue queries can lead to responses either in the form of a structured interface or in the form of general mobile search, when a request is 'out of scope' of the search dialogue. We defined user satisfaction with search dialogues in the generalized form, which showed understanding the nature of user satisfaction as an aggregation of satisfaction with all dialogue's tasks and not as a satisfaction with all dialogue's queries separately. The introduction of dialogue types and understanding which kinds of responses to queries exist helped us to set up a user study and make feature selection for answering the next research question.

Next we investigated **RQ2:** *How can we predict user satisfaction with search dialogues using interaction signals?* To predict user satisfaction, we used the following kinds of interactions: clicks (or 'taps' in terms of touches on mobile platforms), other touch interactions and voice features. The baseline was predicting user satisfaction using clicks and queries features. We showed that features derived from voice and especially from touch interactions add significant gain in accuracy over the baseline. To understand how to efficiently select features depending on different types of queries, we proposed three techniques: using only features of queries resulting in structured interface; calculating a single set of features for queries resulting in structured interface and queries resulting in general SERP; and calculating separate sets of features for each group of queries resulting in structured interface and queries resulting in general SERP. We showed that the third technique is the most accurate to model user satisfaction. This technique improves accuracy from 71% to 81% over the baseline.

Finally, we analyzed the prediction quality of the classifier trained on various selections of interaction features, answering **RQ3:** *Which interaction signals have the highest impact on predicting user satisfaction with search dialogues?* We conducted a feature analysis and concluded that users expect to find answers on the SERP directly without putting in any 'additional effort' (e.g. scrolling). Our analysis showed a strong negative correlation between user satisfaction and swipe actions. Additionally, we demonstrated that users are not satisfied if the intelligent assistant cannot answer their query explicitly and redirects them to a general mobile SERP.

Our general conclusion is that touch based features dramatically improve the prediction quality of user satisfaction with search di-

alogue. Research on intelligent assistants on mobile devices is a new area, and this paper addresses some of the first important and necessary steps. We proposed a method for evaluating user satisfaction with intelligent assistants which can be applied in online evaluation of ranking results, offline mining of user dissatisfaction and understanding directions for their future development.

Acknowledgments

We thank the help from Sarvesh Nagpal and Toby Walker for the help in collecting the internal API data for the user study.

REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *SIGIR*, pp. 345–354, 2011.

[2] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pp. 19–26, 2006.

[3] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *SIGIR*, pp. 773–774, 2007.

[4] E. Bakshy and D. Eckles. Uncertainty in online experiments with dependent data: an evaluation of bootstrap methods. In *KDD*, pp. 1303–1311, 2013.

[5] P. Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf. Res. (IRES)*, 8(3), 2003.

[6] D. J. Campbell. Task complexity: A review and analysis. *The Academy of Management Review*, 1(13):40 – 52, 1988.

[7] A. Deng, T. Li, and Y. Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW*, pp. 609–618, 2014.

[8] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, and C. L. F. Diaz. Towards recency ranking in web search. In *WSDM*, pp. 11–20, 2010.

[9] A. Drutsa, G. Gusev, and P. Serdyukov. Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *WSDM*, pp. 27–36, 2015.

[10] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW*, pp. 256–266, 2015.

[11] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM*, pp. 173–182, 2013.

[12] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR*, pp. 34–41, 2010.

[13] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2):147–168, 2005.

[14] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[15] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Towards estimating web search result relevance from touch interactions on mobile devices. In *CHI Extended Abstracts 2013*, pp. 1821–1826, 2013.

[16] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *SIGIR*, pp. 153–162, 2013.

[17] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM*, pp. 221–230, 2010.

[18] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM*, pp. 2019–2028, 2013.

[19] A. Hassan Awadallah, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: disambiguating long search sessions. In *WSDM*, pp. 53–62, 2014.

[20] L. P. Heck, D. Hakkani-Tür, M. Chinthakunta, G. Tür, R. Iyer, P. Parthasarathy, L. Stifelman, E. Shriberg, and A. Fidler. Multi-modal conversational search and browse. In *SLAM@INTERSPEECH*, pp. 96–101, 2013.

[21] J. Huang and A. Diriye. Web user interaction mining from touch enabled mobile devices. In *HCIR Workshop*, 2012.

[22] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[23] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *ECIR*, pp. 4–15, 2008.

[24] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. G. Kulkarni, and O. Z. Khan. Automatic online evaluation of intelligent assistants. In *WWW*, pp. 506–516, 2015.

[25] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pp. 133–142, 2002.

[26] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pp. 154–161, 2005.

[27] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *FTIR*, 3(1-2):1–224, 2009.

[28] D. Kelly. When effort exceeds expectations: A theory of search task difficulty (keynote). In *SCST@ECIR*, 2015.

[29] D. Kelly, J. Arguello, A. Edwards, and W. ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *ICTIR*, pp. 101–110, 2015.

[30] J. Kiseleva, E. Crestan, R. Brigo, and R. Dittel. Modelling and detecting changes in user satisfaction. In *CIKM*, pp. 1449–1458, 2014.

[31] J. Kiseleva, J. Kamps, V. Nikulin, and N. Makarov. Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them? In *CIKM*, pp. 1561–1570, 2015.

[32] J. Kiseleva, K. Williams, J. Jiang, A. H. Awadallah, I. Zitouni, A. Crook, and T. Anastasakos. Understanding user satisfaction with intelligent assistants. In *CHIIR*, pp. 121–130, 2016.

[33] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD*, pp. 1857–1866, 2014.

[34] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR*, pp. 113–122, 2014.

[35] J. Lehmann, M. Lalmas, G. Dupret, and R. A. Baeza-Yates. Online multitasking and user engagement. In *CIKM*, pp. 519–528, 2013.

[36] J. Li, S. B. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR*, pp. 43–50, 2009.

[37] M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *CSUR*, 34(1):90–169, 2002.

[38] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna. Quality estimation for automatic speech recognition. In *COLING*, 2014.

[39] L. Philips. Hanging on the metaphone. *Computer Language*, 7(12): 39–44, 1990.

[40] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *J AM SOC INF SCI TEC*, 26: 321–343, 1975.

[41] T. Saracevic, P. B. Kantor, A. Y. Chamis, and D. Trivison. A study of information seeking and retrieving. I. background and methodology. II. users, questions and effectiveness. III. searchers, searches, overlap. *J AM SOC INF SCI TEC*, 39:161–176; 177–196; 197–216, 1988.

[42] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR*, pp. 695–704, 2015.

[43] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *WWW*, 2013.

[44] Y. Song, X. Shi, R. White, and A. H. Awadallah. Context-aware web search abandonment prediction. In *SIGIR*, pp. 93–102, 2014.

[45] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD*, pp. 17–26, Washington, DC, 2010.

[46] G. Tür. Extending boosting for large scale spoken language understanding. *Machine Learning (ML)*, 69(1):55–74, 2007.

[47] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Techware: Spoken language understanding resources [best of the web]. *IEEE Signal Process. Mag. (SPM)*, 30(3):187–189, 2013.

[48] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Understanding spoken language. *Computing Handbook*, 3rd ed(41):1–17, 2014.

[49] W. Wahlster. Smartkom: Foundations of multimodal dialogue systems. *Springer*, 2006.

[50] B. Wildemuth, L. Freund, and E. G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 70:1118–1140, 2014.

[51] K. Williams, J. Kiseleva, A. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa. Detecting good abandonment in mobile search. In *WWW*, pp. 495–505, 2016.

[52] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *CIKM*, pp. 91–100, 2014.