Characterizing and Predicting Voice Query Reformulation

Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem and Rosie Jones
Microsoft Redmond, WA USA
{hassanam, raguruna, umuto, rosie.jones}@microsoft.com

ABSTRACT
Voice interactions are becoming more prevalent as the usage of voice search and intelligent assistants gains more popularity. Users frequently reformulate their requests in hope of getting better results either because the system was unable to recognize what they said or because it was able to recognize it but was unable to return the desired response. Query reformulation has been extensively studied in the context of text input. Many of the characteristics studied in the context of text query reformulation are potentially useful for voice query reformulation. However, voice query reformulation has its unique characteristics in terms of the reasons that lead users to reformulating their queries and how they reformulate them. In this paper, we study the problem of voice query reformulation. We perform a large scale human annotation study to collect thousands of labeled instances of voice reformulation and non-reformulation query pairs. We use this data to compare and contrast characteristics of reformulation and non-reformulation queries over a large a number of dimensions. We then train classifiers to distinguish between reformulation and non-reformulation query pairs and to predict the rationale behind reformulation. We demonstrate through experiments with the human labeled data that our classifiers achieve good performance in both tasks.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *selection process, search process.*

Keywords
Voice search, intelligent assistants, voice reformulation

1. INTRODUCTION
Voice-enabled interfaces are becoming a prevalent feature on both mobile and desktop devices. A voice-enabled interface makes human interaction with devices possible through issuing voice commands to the device. Voice-enabled interfaces are used in many scenarios including voice search and interaction with intelligent assistants. While voice interactions are useful and attractive for users, they pose new challenges in terms of understanding how they affect user behaviors that are widely understood with respect to text interfaces.

One such behavior that has been extensively studied in the context of Web search with text input is query reformulation [14]. Query reformulation is the act of submitting a query Q2 to modify a previous query Q1 in hope of getting better results [12]. When searchers experience difficulty in finding information, their struggle may be evident in their search behavior via different indicators such as query reformulation.

Query reformulation has been extensively studied in the literature from different angles including suggesting query reformulation, proposing taxonomies of query reformulation, proposing heuristics and regular expressions to identify different types of reformulation and automatically predicting query reformulation from log data. All this work has been focused on text input, and voice query reformulation has received little attention. Even though many of the studies performed on text queries can be useful for characterizing and predicting voice query reformulation, voice query reformulation has its unique characteristics in terms of the reasons that lead users to reformulate and how the queries are reformulated.

The end to end process of starting from user's search intent to the results that satisfy the user occasionally takes multiple search queries, and the entire process can be seen as a noisy channel problem. In text reformulations, the noisy channel tends to be the fingers -misspelling the search query- and sometimes cognitive errors, when the user cannot phrase the search intent well. Information retrieval literature is rich in spelling and query rewriting techniques to handle these. On the other hand, voice reformulations are different in nature, since the noisy channel is much more in the phonetic space, around the mistakes that the speech recognizer might typically make. This leads to reformulations that would rarely ever happen with text input, such as "ten issues"→"tennis shoes" or "youtube"→"U2".

In this paper, we present and evaluate methods to automatically identify voice query reformulation. Our focus is to characterize the difference between query pairs with and without reformulation. We study different signals that have been used before for predicting text query reformulation and show that many of them are useful for voice query reformulation. We also study several signals that are unique to voice query reformulation and show how they can be used to significantly improve the query reformulation prediction. We make the following contributions with this research:

- Characterize differences in the behavior associated with pairs of voice queries with and without reformulation including voice specific signals such as acoustic signals and statistics of the speech recognition process.
- Perform a large human annotation study to collect thousands of labeled instances of reformulation and non-reformulation query pairs.
- Develop methods to automatically identify voice query reformulations
- Develop methods to automatically identify some of the rationales that lead to voice query reformulation (speech recognition and relevance problems).

The remainder of the paper is structured as follows. In Section 2, we describe related work in query reformulation, query satisfaction and voice interactions. Section 3 describes the labeled data that we use in our analysis. We compare and contrast

CIKM'15, October 19 – 23, 2015, Melbourne, VIC, Australia.
 ISBN 978-1-4503-3794-6/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2806416.2806491.

different characteristics of reformulation and non-reformulation query pairs in section 4. We describe our predictive models in Section 5. Our experiments and findings are summarized in Section 6 and we conclude in Section 7.

2. RELATED WORK

There are three areas of work related to the research presented in this paper: (i) query reformulation, (ii) search satisfaction, success, and frustration, (iii) voice search and intelligent assistants. We now describe previous work in each area in detail and discuss how our method and study extend this prior work.

Query Reformulation: Existing research has studied both how web search engines propose reformulations, and how people reformulate their own queries. Most of the research on the latter has focused on building taxonomies of query reformulation by examining a small set of query logs. Anick [2] classified a random sample of 100 reformulations by hand into eleven categories. Jensen et al. [17] identified 6 different kinds of reformulation states (New, Assistance, Content Change, Generalization, Reformulation, and Specialization) and provided heuristics for identifying them. They used the heuristics to predict when a user is most receptive to automatic query suggestions. The same categories were used in several other studies [10][22]. Huang and Efthimis [14] proposed another reformulation taxonomy. Their taxonomy was lexical in nature (e.g., word reorder, adding words, removing words, etc.). They also proposed the use of regular expressions to identify them. While studying re-finding behavior, Teevan et al. [27] constructed a taxonomy of query re-finding by manually examining query logs, and implemented algorithms to identify repeat queries, equal click queries and overlapping click queries. Another line of research focused on identifying reformulations using classifiers [12], heuristics [14] or regular expressions [17].

Another related problem focuses on the classification of the boundaries of the user search tasks within sessions in web search logs. Boldi et al. [5] presented the concept of the query-flow graph that represents chains of related queries in query logs. Arlitt [3] found session boundaries using a calculated timeout threshold. Murray et al. [23] extended this work by using hierarchical clustering to detect session boundaries. Jones and Klinkner [21] also addressed the problem of classifying the boundaries of the goals and missions in search logs. Lucchese et al. [23] used a similar set of features as [21], but uses clustering to group queries in the same task together. In this work, we extend previous work in that we focus on voice queries by understanding voice query reformulations and studying how we can build models for predicting the relation between pairs of voice queries.

Search Satisfaction: Extensive literature exists on deriving indicators of task success or failure from online user behavior. Fox et al. [8] used an instrumented browser to determine whether there was an association between explicit ratings of user satisfaction and implicit measures of user interest and identified the measures that were most strongly associated with user satisfaction. They found that there was a link between user activity and satisfaction ratings, and that clickthrough, dwell time, and session termination activity combined to make good predictors of satisfaction for Web pages. For example, they found out that a short dwell time is an indicator of dissatisfaction, while long dwell time is correlated more with satisfaction. Behavioral patterns were also used to predict user satisfaction for search sessions. Hassan et al. [20] developed models of user behavior to accurately estimate search success on a session level, independent of the relevance of documents retrieved by the search engine. Ageev et al. [1] proposed a formalization of different types of success for informational search, and presented a scalable game-like infrastructure for crowdsourcing search behavior studies, specifically targeted towards capturing and evaluating successful search strategies on informational tasks with known intent.. Feild et al. [7] developed methods to predict user frustration. They assigned users difficult information seeking tasks and monitored their degree of frustration via query logs and physical sensors.

Our work is different from this work in that we focus on voice-activated interactions with search systems and intelligent assistants. We also do not focus on satisfaction, rather on reformulation which can then be used as an important signal for modeling satisfaction.

Voice Search and Intelligent Assistants: The research direction that's most related to our work is studying characteristics of voice search and its differences from text search. Jiang et al. presents results of a lab study with 20 participants to study query reformulation patterns with voice search, and reports that speech recognition errors significantly degrades performance and users tend to reformulate their queries by removing and substituting words [19]. In their follow up study, Jeng et al. reports that users usually find it difficult to use voice search due to misrecognitions resulting from various reasons [18]. Shokouhi et al. study query reformulations in mobile search, and compare them to reformulations in desktop search [26]. They show that users are less likely to switch between voice and text input methods unless they are starting a new task or correcting recognition errors. Our work extends this work in that it automatically extracts a large set of speech and acoustic features and uses them to build predictive models of voice query reformulation.

There has been significant amount of research on spoken dialog systems in the last two decades [28] and recently partially observable Markov decision processes (POMDP) have become the most commonly used method for managing dialogues [30]. In addition to speech input, intelligent assistants support other forms of input such as typing, or selecting displayed results, therefore regarded as multi-model conversational systems [29]. Recently, Jiang et al. present an automatic method from online data to evaluate intelligent assistants [13]. They build models to evaluate overall user satisfaction, as well as models to evaluate speech recognition and intent classification independently, and they also identify various behavioral patterns that relate to user satisfaction.

Contributions of Our Study: We extend previous research in a number of ways. First, we devise and evaluate methods to automatically identify voice query reformulation. Second, in contrast to prior work on query reformulation, we not only use text based features but also look at other signals unique to voice interactions like phonetic similarity, change in speaker emotions, speech rate, etc. Finally, we focus on identifying the rationales that lead the user to trying to reformulate the request providing useful insight that can help system developers improve the system performance and hence improve the overall user experience.

3. DATA & ANNOTATION GUIDELINES

In this section we describe the data we use in this study. We also describe the human annotation study we conducted to collect labels for this data.

3.1 Data

Our data consists of a sample of query pairs from the user logs of a major search engine. The data was sampled given the following conditions:

- Q1 is a voice query, Q2 can be of any input type (i.e. voice, text, autosuggest)
- The time difference between Q1 and Q2 is less than 5 minutes
- Q1 has received no clicks

The first condition is intended to focus on voice queries which are the main focus of this study. The remaining two conditions are intended to generate a high-recall low-precision filter that retrieves most of the reformulation pairs. This is following the work presented in [12] which showed that most query reformulations occur within 5 minutes of the original query and that the original query does not usually receive any clicks. Note that if we do not apply such filters before sampling, most of the sampled query pairs will not have a reformulation limiting our ability to study voice query reformulation [12]. Also note that once we build the query reformulation predictor, as will be described later, we can apply it freely to the logs without the need to apply such filters.

We randomly sampled such query pairs from user logs of a major search engine between June and December 2014. In order to remove variability caused by geographic and linguistic variation in search behavior, we only included queries generated in the English speaking United States locale. The query pairs were then labeled by human annotators. Annotators were recruited from the crowdsourcing service Clickworker.com, which provided access to crowd workers under contract. Annotators resided in the United States and were fluent in English. We collected human annotations on these query pairs along different dimensions: (1) speech recognition quality, (2) relevance of the results of Q1, and (3) whether Q2 is a reformulation, a related query, or an unrelated query. Each query pair was judged 5 times independently by different annotators and the majority label was used. We defined reformulation, related queries and unrelated queries as follows:

Definition: A *reformulation* is a query pair where the user is submitting a query Q2 to modify a query Q1 while still trying to satisfy the same information need as in Q1.

In voice queries, users opt to reformulate their queries if they think the results do not satisfy their original information need either due to a problem with the speech recognition or due to the retrieved results being not relevant. When users decide to reformulate their queries, they may issue another voice query with the exact words but pronounces it differently or speaks at a different pace. The user may also decide to type in what she said in Q1, fixing speech recognition errors, for example "watson's ministy"→"watson's manistee". Additionally, the user may decide to use different wording without changing the meaning, such as "pictures of kane's flowers"→"cannis flowers photos".

Definition: A *related query* is a query Q2 that has the same topic as Q1, but is not a reformulation of Q1.

Related queries are intended to satisfy a different information need and hence are not considered as a reformulation. Examples of related queries include "just dance"→"just dance videos", or "toyota corolla mpg"→"toyota yaris mpg".

Definition: An *unrelated query* is a query Q2 that is neither related to nor a reformulation of Q1.

3.2 Annotation Guidelines

The annotation task consists of listening to Q1 and examining the speech recognizer output for this voice query (and the same for Q2, if Q2 is also a voice query), and inspecting search result pages for both these queries, then answering three questions. To avoid overloading the annotator with all the information, the instructions are interleaved with the relevant questions, but the judge is encouraged to come back to the earlier questions at any time. Also the annotators are allowed to listen to the voice queries as many times as they like. In addition, to provide more context to the annotators, they were given the following information: the timestamps of each query, the search result pages of each query, the clicks the second query received and the fact that the first query received no clicks. The annotation process proceeded as follows:

1. The judge listens to audio of Q1 and examines the speech recognizer output for this query and answers Question 1 (SR quality): *Was Q1 recognized correctly*?
   PERFECT: if all the words are recognized correctly.
   GOOD: minor errors only, a word or two missed or misrecognized, but overall content is pretty well covered.
   BAD: a key word is missing or misrecognized
2. The judge is instructed to click on the recognized text of Q1, inspect search results and answer the second question. The judge can return and change the answer after listening to (or looking at) Q2: Question 2 (search relevance): *Did the user find what were they looking for in Q1?*
   PERFECT: the user found the required information
   GOOD: the user found some useful/relevant information
   BAD: the user did not find what he/she was looking for
3. The judge listens to Q2, and inspects the search results for Q2 and answers Question 3: *Is Q2 a reformulation of Q1?*
   YES: Q2 is a reformulation of Q1. Here note that this may be a reformulation, even if the recognized text for Q1 and/or Q2 do not correctly represent what the user said
   RELATED: Q2 is related to Q1, but not exactly the same information need
   NO: User is looking for something new

Our sample originally contained 7500 query pairs. After excluding query pairs that judges indicated have a problem with understanding the audio and they cannot provide a judgment, we ended up with 7322 labeled query pairs. When manually investigating these cases, we found out that these are cases where the audio is not intended for the speech recognizer, such as pocket dialing, or a child playing with the phone saying nothing intelligible.

4. CHARACTERIZATION OF VOICE REFOMULATION BEHAVIOR

In this section, we examine several characteristics of voice reformulation focusing on query similarity, speech quality, result quality and acoustic dimensions.

4.1 Query Similarity

At the outset of our analysis, we examine a number of different ways of measuring the similarity between a pair of queries and compare reformulation pairs to non-reformulation pairs. We might expect non-reformulation pairs to contain less overlap as people revise their queries to explore alternatives. We use three different methods for assessing the similarity between two queries: (1) lexical similarity, (2) semantic similarity, and (3) phonetic similarity.

Lexical Similarity: To measure the lexical similarity between pairs of queries, we begin by performing standard text normalization where we lowercase the query text, and remove stop words. Then we represent every query as a bag of non-stop word terms. The similarity between any two queries Qi and Qj is computed as follows:

$$\frac{|Qi \cap Qj|}{|Qi| + |Qj| - |Qi \cap Qj|}$$

where $|Qi|$ is the number of terms in query Qi, and $|Qi \cap Qj|$ is the number of matched terms in Qi and Qj. To calculate the number of matches in Qi and Qj ($Qi \cap Qj$), we consider two terms matched if: (1) the two terms match exactly, (2) to capture spelling variants and misspelling, we allow two terms to match if the Levenshtein edit distance between them is less than two, or (3) We match two terms if the lemmas of their tokens match. Lemmatization is the process of reducing an inflected spelling to its lexical root or lemma form.

Semantic Similarity: Latent semantic models, such as LSA, map text to a low-dimensional space where the similarity between two pieces of text is easily computed by measuring the distance between the vector representations of the two pieces of text. Recent developments in this area have used neural networks to learn vector representation of text, e.g. [24][15]. One use of these representations is to find semantic similarity between texts beyond what lexical bag-of-words representation can accomplish.

We used the model described in [15], to map each query to a vector space representation. The similarity between any two queries is then computed by measuring the cosine similarity of the two vectors

Phonetic Similarity: Voice query reformulation has unique characteristics compared to text query reformulation. There are multiple ways users reformulate their text queries [14]. One of the common ways people reformulate text queries is by introducing some lexical variations, e.g. "knoledge management" → "knowledge management", or a semantic variation, e.g. "barber" → "hair cut". In addition to these sources of errors, voice queries have a unique source of errors that result from mistakes made by the automatic speech recognition system, e.g. "u2" → "youtube". These cases cannot be identified using the lexical or semantic similarity measures and require special handling.

Philips [25] introduced an algorithm for reducing English words to their basic sounds. This allows us to index words by their pronunciation and hence identify query reformulations resulting from speech recognition errors. For example, a query "WhatsApp" may be incorrectly recognized as "what's up". Both lexical and semantic similarity will not find these two texts similar. However, the Metaphone code for both of them is "WTSP". In such cases, phonetic similarity can help identify query reformulations that happen in response to speech recognition errors. We transform the transcribed text of each query into Metaphone codes using [25]. We then compute the normalized Levenstein edit distance between the Metaphone representations to estimate the phonetic similarity.

In Figures Figure 1, Figure 2, and Figure 3, we plotted the distribution of the lexical, semantic and phonetic similarities for query pairs with and without reformulations. In all figures, the x-axis represent similarity values bins and the y-axis represent the percentages of query pairs in each bin. Since lexical and phonetic similarities are based on normalized Levenstein edit distances, their values range between 0 and 1 with 0 indicating a perfect match and 1 indicating no match at all. Since the semantic similarity is based on the cosine similarity between two vectors, it ranges between -1 and 1 denoting the weakest and the strongest similarity respectively. Note that the measure of lexical and phonetic similarity is a distance measure (i.e. the smaller, the more similar) while the measure of semantic similarity is a similarity measure (i.e. the larger, the more similar).

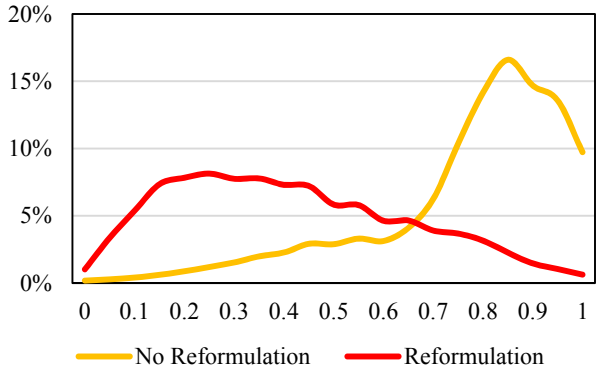

Figure 1. Lexical similarity in query pairs with and without reformulation

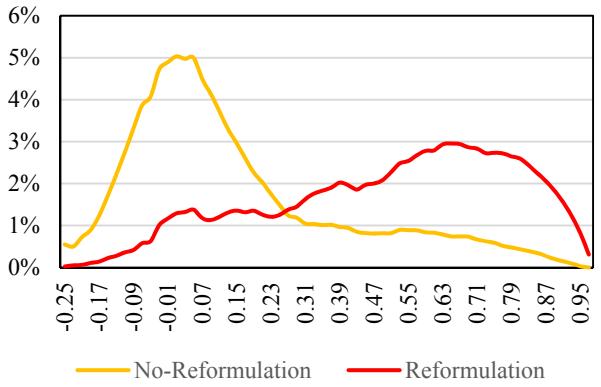

Figure 2. Semantic similarity in query pairs with and without reformulation

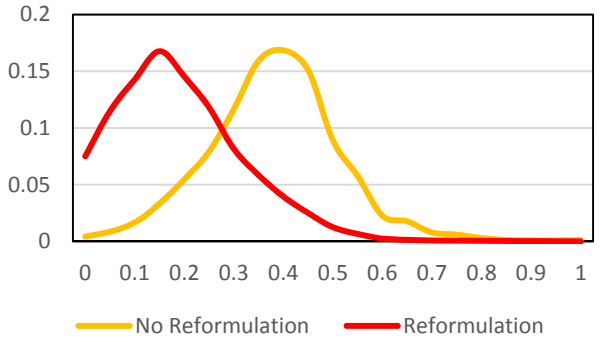

Figure 3. Phonetic similarity in query pairs with and without reformulation

Looking at the three figures, we notice that the distribution of all types of similarities is significantly different in the case of reformulation as opposed to the case of no reformulation. In all three cases, reformulation pairs had higher similarity/less distance. Note that the difference between the means for all three types of similarity with spastically significant ($p < 0.001$) using a two-tailed t-test. We will show later that the three types of similarities are also complementary in covering different types of query reformulation.

4.2 Speech recognition Quality

We now turn our attention to speech recognition quality with the objective of studying how it affects reformulations. We start by

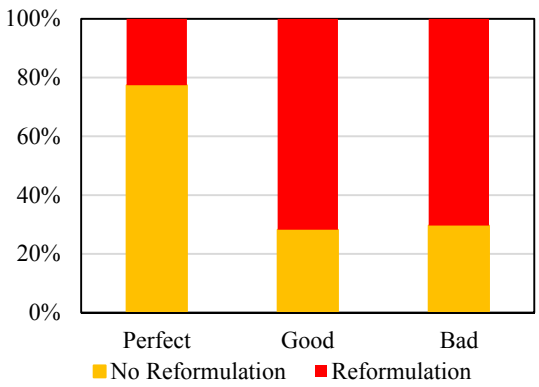

Figure 4. Relation between query reformulation and speech recognition quality as labeled by annotators

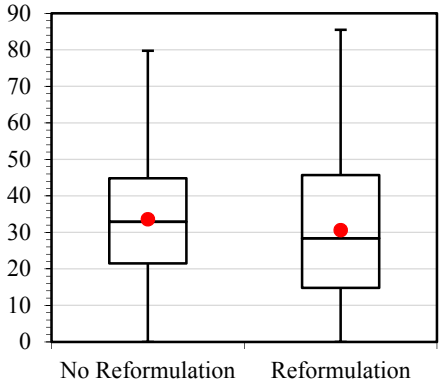

Figure 5. Box-and-whisker plot for speech recognition confidence for reformulation and no-reformulation cases. Mean is dot. Median is horizontal line.

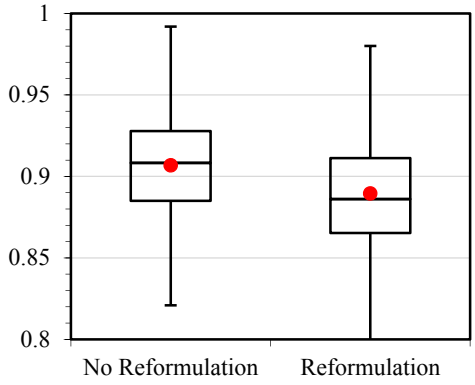

Figure 6. Box-and-whisker plot for language model score for reformulation and no-reformulation cases. Mean is dot. Median is horizontal line.

looking at speech recognition quality labels. Recall that we asked our annotator to judge the speech recognition quality on a 3-point scale (Perfect, Good, or Bad). We split the query pairs according to the speech recognition quality label of the first query. For every label$x$, we computed the proportion of queries that has been reformulated and the proportion of queries that has not. We show the results in Figure 4. Relation between query reformulation and speech recognition quality as labeled by annotators. We notice that only slightly over 20% of the queries are reformulated when the speech recognition is perfect. This quickly increases to almost 70% when the speech recognition quality is good or bad. This suggest that errors with speech recognition significantly increase the likelihood of the query being reformulated. It also shows that this happens both for minor and major speech recognition errors. Since human labeled assessment of speech recognition quality is available only for the labeled dataset, we now turn our attention to other properties of the speech recognition system that, while noisy, we suspect might be useful for use as a proxy for the speech recognition quality.

The first property we look at is the speech recognition system confidence. Confidence scores of automatic speech recognition system decoding [16]. In Figure 5, we show a box-and-whisker plot of the speech recognition confidence scores of recognizing the first query when it is followed by a reformulation and when it is not followed by a reformulation. The figure shows that queries that end up being reformulated typically have lower speech recognition confidence scores.

In addition to the speech recognition confidence score, we also use the language model score of the recognized text. Higher language model scores indicate higher relative likelihood of the recognized phrases given its context. A Box-and-whisker plot for language model score for reformulation and no-reformulation cases is shown in Figure 6. Like the speech recognition score, we notice that reformulated queries tend to have lower language model scores. The differences between the means for both speech recognition confidence and language model score were statistically significant ($p < 0.001$) using a two-tailed t-test.

4.3 Result Quality

Previous research [12] has shown that reformulation is a strong sign of dissatisfaction with the results of the current query. We asked our annotators to label the relevance of the results to the query on a 3-point scale (Perfect, Good, or Bad). Similar to what we did with the speech recognition quality labels, we split the query pairs according to the result quality label of the first query. For every label$x$, we compute the proportion of queries that has been reformulated and the proportion of queries that has not. We show the results in Figure 7. Confirming the findings of previous work, we see that the proportion of reformulated queries increasing as the quality of the results decrease showing that bad result quality is another factor, in addition to speech recognition quality, that can lead to query reformulation.

Since many queries can be satisfied by elements in the search result page (e.g. answers) without the need to click on any results, we looked at the time the user spends on the result page before entering the following query. Our hypothesis is that the time between queries in case of reformulation will be shorter than in case on no reformulation. To validate this hypothesis, we plot the distribution of time between queries in both reformulation and non-reformulation cases in Figure 8. We notice that it is indeed the case that users tend to spend less time between queries when the query is reformulated. This probably happens because the users quickly realize that the current results do not meet their needs and hence they decide to reformulate the query. On the other hand, when the query is not reformulated, the user may spend more time examining elements of the results page that may satisfy her information need (e.g. answers). This agrees with previous results reported on conventional (non-voice) search queries [12].

Finally, we look at the clicks received by the second query in cases where the second query was or was not a reformulation of the first query in Figure 9. Recall that our dataset was sampled

such that all first queries have received no clicks. Hence, we only look at the second queries when studying clicks. Figure 9 shows that when the second query received a click, it is more likely to be a reformulation of the previous query. This probably happens because users find what they are looking for after reformulation.

4.4 Summary

In this section, we have characterized some key aspects of voice query reformulation. We have shown that there are several key differences between reformulation and non-reformulation queries. We have shown that previously studied query similarity measures, i.e. lexical similarity, is also effective for voice queries. We have also shown that other semantic similarity measures and similarity measures specific to voice queries, i.e. phonetic similarity, are also very effective. Additionally, we have shown that there are key characteristics related to the system performance, namely speech recognition quality and results quality, that play an important role in determining whether a query will be reformulated or not. Since these characteristics require human labeled data and are not readily available at production time, we have also shown that there are other system and user behavior signals that can be used as proxies for these properties and that also do differ between reformulation and non-reformulation queries. In the next section, we describe a model that leverages these features and others to distinguish between reformulation and non-reformulation query pairs.

5. REFORMULATION PREDICTION

In this section, we formally define our prediction task and then introduce the features used for prediction.

5.1 Problem Definition

As we stated earlier, voice Query Reformulation is the act of submitting a query Q2 to modify or repeat a previous search query Q1 in hope of avoiding misrecognition or retrieving better results. Hence, query reformulation is considered an indication of dissatisfaction with the previous query. For Q2 to be considered a reformulation of Q1, both queries must be intended to satisfy the same information need. A query reformulation happens when the user is not satisfied with the recognition of her query (e.g. speech recognition error), the results of the query or both.

Note that a related query on the same topic addressing a different information need is *not* considered as query reformulation for our purpose. Given a pair of queries, our objective is to predict whether the second query is a reformulation of the first query or not. For each query pair, we have the audio and recognized text of the first query and audio and recognized text of the second query if it was a voice query. Otherwise we have the text of the second query. Note that the first query is always a voice query, while the second could be either a voice query or a text query. We also have any clicks made by the user and the timestamps of all queries and clicks. Our second objective is to predict the rationale behind reformulation (speech recognition error or relevance issue) for queries that are known to be reformulated.

5.2 Features

Our predictive model uses the following groups of features:

Session features: These are the features that describe general characteristics of the current session such as the length of the two queries, the time between them, clicks etc. They also cover the input method and more particular any switches in the input method for voice to text. Note that the data is sampled such that all first queries are voice queries. Hence, we only used the input method of the second query as a feature.

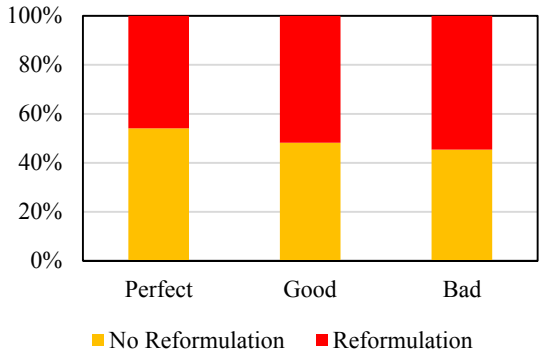

Figure 7. Relation between the relevance of the first query results and reformulation

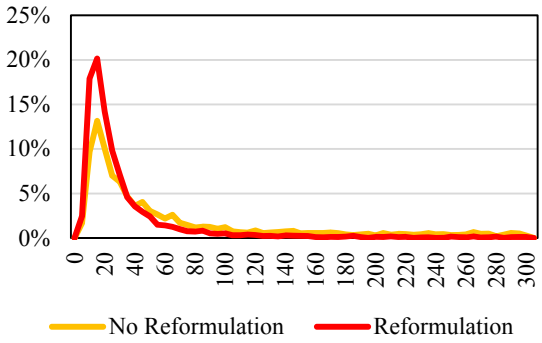

Figure 8. Time between queries for instances with and without reformulation

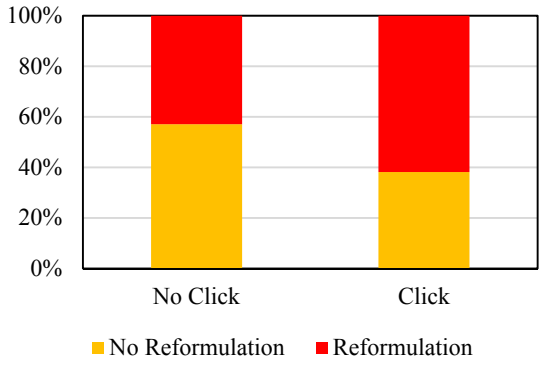

Figure 9. Relation between clicks on the second query

Lexical similarity features: The second group of features characterizes the lexical similarity between the two queries. Previous research, e.g. [21][12], has shown that reformulation query pairs are more likely to have higher lexical similarity than other consecutive query pairs. We characterize lexical similarity using a number of features including number of words in common, Jaccard coefficient between a bag of words representation of the two queries, whether the first query has been generalized (1+ terms are removed from the previous query) and whether the first query has been specified (1+ terms are added to the previous query). We also use a lexical similarity formula that combines edit distance and words in common as described in Section 4.1.

Phrase similarity features: As discussed earlier lexical similarity typically treats queries as a bag of words ignoring any structure in the text. Following previous work on text query reformulation [12], we introduce a number of features that take the phrase structure into consideration. Instead of the bag of words representation, queries are represented as a bag of keywords. For example, the query "weather in San Francisco" contains two keywords "weather" and "San Francisco". To build such a representation, we segment each query into keywords. Query segmentation is the process of taking a user's search query and dividing the tokens into individual phrases or semantic units [3]. Consider a query $x = \{x_1, x_2, \dots x_n\}$ consisting of n tokens. Query segmentation is the process of finding a mapping: $x \rightarrow y \in Y_n$, where y is a segmentation from the set Y_n. There are many approaches to query segmentation that have been presented in the literature. Some of them are supervised, e.g. [3], while others are unsupervised, e.g. [11]. We opt for using the unsupervised method presented in [11] especially since it has been shown it achieves similar performance to those that are trained on large labeled datasets. Once the phrase representation built, we extract a number of features similar to the lexical features including the total number of keywords in each query, and number of keywords in query 1/2 only. We also use the same similarity function that takes both edit distance and word overlap into consideration but applied to keywords rather than individual words.

Semantic similarity features: Lexical and phrase features cover the cases where the query is reformulated to add/remove terms or introduce some lexical variation to the original query. There are cases though where the user decides to use different terms, albeit semantically equivalent, to reformulate the query. We use the vector space representation described in Section 4.1 to represent each query as a vector and then use the cosine similarity to assess the semantic similarity between two queries by measure the cosine of the angle between their vector space representations.

Speech Features: Voice queries have the unique characteristic that the automatic speech recognition step adds another layer of noise between what the user says and what the system recognizes. This introduces a new reformulation pattern that arises from mistakes by the speech recognition engine. For example the speech recognition engine can easily mistake a search for the band "u2" as a search for the video sharing website "youtube". To capture such cases, we represent each query with the Metaphone representation and assess the *phonetic similarity* by the edit distance between the Metaphone representations of the queries.

We also add features that are motivated by our finding in Section 4.2 where we showed that reformulation is strongly dependent on the *speech recognition quality* as labeled by human judges. Since it is not feasible to have human judges label each voice query for its speech recognition quality, we instead use statistics of the speech recognition system that can act as a surrogate for the recognition quality. More specifically, we use the speech recognition confidence, the language model confidence and the recognition time. Since the speech recognition engine typically generates many candidate recognitions and then selects the most likely one, we also introduce features to assess the similarity between the reformulated query and the n best candidates generated by the speech recognition engine for the first query.

Finally, we add features to describe the acoustic characteristic of the utterances of the voice queries. We hypothesize that features of the acoustic signals of the utterance, as well as features characterizing the change in the acoustics from the first to the second query can be very useful for detecting query reformulation since users who select to reformulate their queries are likely to have a change in speed, loudness, pitch etc. as a way of showing frustration or partially emphasizing the words that where misrecognized. Partial emphasis, as labeled by human judges, have been shown earlier to correlate with reformulation [19]. Here, we automatically derive acoustic features that can act as a proxy for frustration, partial emphasis and other behaviors. To generate the pitch and loudness features, we use the openSmile toolkit [6]. openSmile is an open source large scale multimedia feature extraction toolkit that allows users to easily extract features from both audio and video files.

Table 1. Features used to distinguish between reformulation and non-reformulation query pairs. Features marked with an "*" were computed for both the first and the second query

Name	Description
Session Features	
CharQueryLen*	Query length in number of characters
WordQueryLen*	Query length in number of words
TimebetQueries	Time between queries
Query2Clciked	True if Q2 received any clicks
Voice2Text	True if user switched from voice input in Q1 to text input in Q2
Lexical Query Similarity Features	
AvgQuerySim	Similarity between queries (see Section 4.1)
WordsInCommon	Number of words in common
JaccardDistance	Jaccard distance between queries
Query2IsGen	1 if the second query has 1+ terms are removed from the previous query
Query2IsSpec	1 if the second query has 1+ terms are added to the previous query
Phrase/Keyword Similarity Features	
NumKWMatch	Number of keywords that match in both queries
NumKWinQuery*	Number of keywords in query
KWinQ1Only	Number of keywords in Q1 but not Q2
KWinQ2Only	Number of keywords in Q2 but not Q1
Semantic Similarity Features	
SemanticSim	Cosine similarity between semantic vector representations of the queries
Speech Features	
PhoneticSim	Levenstein distance between Metaphone representation of the two queries
SRConf	Speech recognition confidence
LMScore	Language model score
SRRecoTime	Speech recognition time
Q2NBestExactMatch	Q2 is in N best candidate recognitions of Q1 (exact match)
Q2NBestPartialMatch	Q2 is in N best candidate recognitions of Q1 (partial match)
DeltaWordRate	Diff. in speech rate (words/sec)
DeltaLoudness	Diff. in voice loudness
DeltaPitch	Diff. in voice pitch

5.3 Prediction Tasks

Using the features presented in the previous section, we build predictive models to tackle two main tasks. The first task is given a pair of queries, predict whether the second query is a reformulation of the first query or not. The second task focuses only on reformulation queries. There are multiple reasons that lead users to reformulate their queries. Two of the main reasons are speech recognition mistakes and relevance issues. To study this further, we collected additional labels, during the human annotation study described in Section 3, of whether there was a

Table 2. Voice reformulation prediction performance. * and + indicate statistical significance at $p < 0.05$ using paired t-tests compared to the B1 and B2 respectively. A majority baseline has a 53% accuracy.

		Accuracy	*Precision*	*Recall*	*F1*
1	**Lexical/Phrase**	72.93%	77.56%	69.65%	73.39%
2	**Semantic**	77.07%	79.34%	77.36%	78.34%
3	**Session**	64.47%	64.84%	73.63%	68.96%
4	**Speech**	79.47%	85.13%	74.75%	79.60%
5	**B1: Lexical/Phrase**	72.93%	77.56%	69.65%	73.39%
6	**B1 + Sem.**	81.27%*	82.48%*	82.59%*	82.53%*
7	**B1 + Speech**	85.13%*	85.91%*	86.44%*	86.17%*
8	**B2: B1+ Session**	81.67%	83.35%	82.21%	82.78%
9	**B2+ Semantic**	82.47%+	84.03%+	83.08%+	83.55%+
10	**B2 + Speech**	84.53%+	83.66%+	85.25%+	84.45%+
11	**All**	**86.27%**	**87.28%**	**87.6%**	**87.45%**

Table 3. Voice reformulation rationales (speech recognition and relevance issues) prediction performance. A majority baseline for speech recognition issues has a 68% accuracy and for relevance issues has a 79% accuracy.

		Accuracy	*Precision*	*Recall*	*F1*
1	**Speech Reco.**	83.96%	86.36%	92.65%	89.40%
2	**Relevance**	87.15%	89.70%	95.64%	92.57%

problem with the speech recognition or not and whether the results of the first query were relevant or not. Since the two issues are not mutually exclusive, we build separate predictive models for each. We discuss the details of these experiments and the experiments conducted to evaluate the reformulation predictive model in the next section.

6. EXPERIMENTS AND EVALUATION

In this section we describe the experimental setup and the results for the prediction experiments. We also include a detailed analysis of the performance of the features that were used in the prediction experiment.

6.1 Experimental Setup

We performed several experiments to study the performance of our predictive models. We conducted experiments using the data described in Section 3 which had 7322 labeled query pairs along with their audio file, if applicable, their timestamps and click information. We trained a classifier to distinguish between reformulation and non-reformulation query pairs. We also trained two other classifiers focusing on reformulated queries to predict whether there were issues with the speech recognition and/or relevance of the results of the first query.

In addition to using all the features we described in Section 5.2, we also experiment with several baselines that have access only to a subset of these features including the state-of-the-art features used for predicting query reformulation using text-input [12]. More specifically, we experimented with the following baselines:

Individual Feature Groups: The first set of baseline uses each feature group from Table 1 individually. This results in the following baselines: Lexical/Phrase, Semantic, Session, and Speech. Note that we treat lexical and phrase feature as one feature group and phonetic similarity and other speech signals as one feature group for simplicity.

Baseline 1 (Lexical/Phrase): This baselines implements all features from the literature that predict query reformulation bases on lexical and phrase similarity between the two queries.

Baseline 2 (Lex/Phrase + Session): This baseline adds session features to the previous baseline.

In addition to the baselines listed above, we experiment with adding each new feature group (semantic, and speech) to both baselines and finally, we show the performance with all features.

We used 10-fold cross validation for all experiments and Multiple Additive Regression Trees (MART) for classification [9].We experimented with other classifier (SVM and logistic regression), and MART yielded better performance.

6.2 Results and Findings

We now present the results of the prediction experiments. Table 2 presents the results of the reformulation prediction experiment. We report the performance of multiple classifiers corresponding to the different feature groupings we introduced in the previous section. For every classifier, we report accuracy, precision, recall and F1 measure. We also compare the classifiers to the majority baseline that always predicts the label of the dominant class.

We notice from the table that using only lexical and phrase features yields much better results than using marginal class distributions (53% accuracy). This shows the predictive power of the lexical and phrase features. This can be explained by the analysis we presented in Section 4.1 which shows that lexical query similarity is a clear distinguishing factor when it comes to telling reformulation query pairs apart from non-reformulation query pairs. This also shows that voice query reformulation is similar in some ways to text query reformulation and that there is a lot of room for improvement if we can find features that characterize the unique nature of voice query reformulation compared to text query reformulation.

We also notice that using semantic similarity features results in a performance that surpasses lexical and phrase similarity features (2nd row). This can also be explained by the analysis presented in Section 4.1 which highlighted the discriminative power of semantic features. Semantic features are particularly useful for cases where the user opts for using different words with the same meaning or even for cases where the user is still using the same words but with lexical variations that the edit distance measure fails to recognize.

As for the session features (3rd row), we notice that they are not as predictive as the similarity features. We also notice that they do a better job, in terms of F1 measure, in identifying the positive class (reformulation) compared to the negative class (no-reformulation). That said, they still yield better results than marginal class distributions (53% accuracy). Although session features are certainly useful, our data was collected in a way that makes the reformulation predicting task harder especially for session features since we restrict it to query pairs with no clicks on the first query.

The last feature group (4th row) contains all features that are unique to voice query reformulation like phonetic similarity, and other acoustic features. We notice from the table that this feature group performs significantly better than all other feature groups. They particular achieve very high precision showing that they can accurately identify reformulation cases. This supports the hypothesis we presented earlier that voice query reformulation

Table 4. Feature group importance.

Reformulation	Speech Reco.	Relevance
Speech	Speech	Semantic
Semantic	Semantic	Lexical
Session	Lexical	Session
Lexical	Phrase	Speech
Phrase	Session	Phrase

has different characteristics that cannot be captured by using session based features or features derived only from the text and ignoring the speech signal.

In the second part of the table (rows 5-7), we pivot on the lexical/phrase features as a baseline and study the effect of adding semantic and speech features to the baseline. Our objective is to understand whether the gains we are getting from the different feature groups are additive or not. Looking at rows 6 and 7, we see that adding both semantic and speech features improves the baseline with speech features resulting in a bigger gain compared to semantic features. All difference are statistically significant at $p < 0.05$ using a paired t-test. This is a desirable outcome as it shows that the gains from adding the semantic and speech features add up to the gains we get from the lexical/phrase baseline.

Moving on to the third part of the table (rows 8-10), we pivot on a second baseline which combines lexical, phrase and session features. These represent features from state-of-the-art work on reformulation prediction [12] and task boundary detection [21]. The table shows that, similar to the first baseline, adding the semantic and speech features to the second baseline also improves the performance. Also, we notice that adding speech feature results in a bigger gain than adding the semantic feature confirming that the speech feature do a better job in identifying voice reformulation. Finally, when we combine all features and train a single classifier (last row), we get a better performance than any other combination of features showing the usefulness of all feature groups in identifying voice reformulation.

We now move on to another set of experiments where we focused on reformulation queries and we tried to predict the rationale behind the reformulation. Recall that in our human annotation study, we collected labels for whether the user has faced any issues with the speech recognition quality and/or result relevance. These labels were collected on a 3 point scale. To build a binary classifier, we assigned the cases where no or few speech recognition errors were detected (and the case where results were relevant/somewhat relevant) to one class and the cases where many speech recognition errors were detected (and the cases where results were not relevant) to another class.

The results of these experiments are shown in Table 3. The table shows that we can predict the rationale behind reformulation with a high accuracy. Note that we do not claim that this is an exhaustive list of all the reasons that lead to query reformulation. Rather, we just focused on two of the main reasons that may lead users to reformulate their voice queries.

We now turn our attention to the features that contribute most to the prediction task. In order to assess the importance of the proposed features, we compute the information gain of all features. We do that by computing the error reduction for each feature at each node split using the squared loss function. We then aggregate the error reduction at all splits for every feature and use that as a measure of feature gain. We do not report feature gains at the individual feature level for space considerations. Instead, we report feature importance at the feature group level for groups in Table 1. Features used to distinguish between reformulation and non-reformulation query pairs. Features marked with an "*" were computed for both the first and the second query. We assigned every group a rank based on the average score of all its features in the top 10 list.

The results are shown in Table 4. It is clear from the table that speech and semantic features play the most important role. For predicting voice reformulation and speech recognition problems followed by the other feature groups. As for the relevance problems prediction, semantic and lexical features are ranked higher than other features. If we dig deeper into the speech feature group, we find out phonetic similarity, speech recognition confidence and acoustic features (change in word rate, loudness and pitch) contribute the most in the same order they were mentioned.

6.3 DISCUSSION AND IMPLICATIONS

Identifying pairs of reformulated queries is a very well-studied topic with multiple applications in text queries. However, this topic has received little attention when it comes to voice queries. A recent study[1], executed by Northstar Research and commissioned by Google, found out that 55% of U.S. teens use voice search every day and that 89% of teens and 85% of adults agree that voice search will be "very common" in the future. This shows the importance of studying and understanding important behaviors such as reformulation in the context of voice search. In this paper, we showed that we can build a classifier to identify voice query reformulation and the rationale behind them accurately. We also show that using speech related features significantly improves the prediction accuracy compared to features used for text queries.

There are at least two potential areas of future work that can increase the impact of the work in this paper. The first is improving the way by which the labels were collected for our data. The labels were assigned by third-party judges based on their consensus opinion of the two queries. While using the consensus may improve the reliability of the label, the fact that the judges were not those who submitted the queries may lead to unforeseen issues with the labels. Ways to address this shortcoming include soliciting judgments from the users in-situ at the time of their interaction with the system.

Second, we focused on how to build predictive models for predicting reformulation and the rationale behind it in this paper, and did not discuss its applications. Our contributions have potential implications on many applications. Detecting pairs of queries reformulated by the user is an important and necessary first step toward addressing the task of automatic query reformulation prediction. Query reformulation can be used to find chains of queries as constructed by the user. These chains provide us with labeled examples of how queries can be rewritten to improve the user experience. Using this labeled data, we can train system to automatically reformulate queries when it predicts that the user is/will not be satisfied with the current results. This has several applications ranging from learning to automatically rewrite queries or providing query suggestions in order to get the user to the correct results faster [2] or segmenting user queries into tasks [21].

[1] http://prn.to/1sfiQRr

Additionally, reformulation is a sign of struggle which has been shown in previous work to be one of the strongest factor behind user dissatisfaction [12]. This suggests that we can use our model for identifying cases where the users are dissatisfied with the system performance. This in turn can be used to generate such dissatisfaction cases that can be used by search systems developers to identify problems with their systems. This can be very useful when coupled with other predictive models that we developed that can decide whether the problem is with the speech recognition quality, the result relevance or both. Alternatively, it could also be used as a metric to compare different systems, and the rate of voice query reformulations can serve as a proxy for the quality of the user experience.

7. CONCLUSIONS

Reformulation is one of the clearest signs of struggle when interacting with a search system or an intelligent assistant. Although query reformulation has been extensively studied before in the context of Web search, voice query reformulation has not received as much attention. Voice query reformulation is similar to text query reformulation in many aspects. However, it also has unique characteristics in terms of the reasons that lead users to reformulation and the features that can be used to identify it. We have shown that there are differences in behavioral attributes such as phonetic query similarity, input method switching, etc., as well as system attributes such as speech recognition confidence, recognition time, etc. between reformulation and non-reformulation query pairs. These differences can be useful in distinguishing reformulation query pairs from other query pairs. We have also developed classifiers and showed that we can perform this prediction accuracy. We have also shown that we can accurately predict the reason behind query reformulation (e.g. speech recognition mistakes, non-relevant results or both). Future work involves the expansion of our research in this area to consider other applications of the predictors such as finding cases of dissatisfaction and helping users recover from such cases.

8. REFERENCES

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

[2] Anick, P. (2003). Using terminological feedback for web search refinement:a log-based study. In Proc. SIGIR, 88–95.

[3] Arlitt, M. (2000). Characterizing Web user sessions. ACM SIGMETRICS Performance Eval Review, 28(2), 50–63.

[4] Bergsma, S., and Wang Q. I. (2007). Learning Noun Phrase Query Segmentation. In Proc. EMNLP, 819–826

[5] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. (2008). The query-flow graph: model and applications. In Proc. CIKM, 609-618.

[6] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proc. ACM'MM, 1459−1462.

[7] Feild, H., Allan, J., and Jones, R. (2010). Predicting searcher frustration. In Proc. SIGIR, 34−41.

[8] Fox, S., Karnawat, K., Mydland, M., Dumais, S.T., and White, T. (2005). Evaluating implicit measures to improve the search experience. ACM TOIS, 23(2), 147−168.

[9] Friedman, J.H., Hastie, T., and Tibshirani, R. (1998). Additive Logistic Regression: A Statistical View of Boosting. Technical Report, Stanford University.

[10] Guo, J., Xu, G., Li, H., and Cheng, X. (2008). A unified and discriminative model for query refinement. In Proc. SIGIR ‘08, 379-386.

[11] Hagen, M., Potthast, M. Stein, B, and Bräutigam, C. (2010). The Power of Naïve Query Segmentation. In Proc. SIGIR, 797–798.

[12] Hassan Awadallah, A., Shi, X., Craswell, N., and Ramsey, B. (2013). Beyond Clicks: Query reformulation as a predictor of search satisfaction. In Proc. CIKM, 2019–2028

[13] Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Kulkarni, R. G., and Khan, O. Z. (2014) Automatic Online Evaluation of Intelligent Assistants. In Proc. WWW.

[14] Huang, J. and Efthimis N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In Proc. of CIKM, 77−86.

[15] Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In Proc. CIKM,2333−2338.

[16] Huang, P.S., Kumar, K., Liu, C., Gong, Y. and Deng, L. (2013). Predicting speech recognition confidence using deep learning with word identity and score features. *Proc. ICASSP*, 7413–7417

[17] Jansen, B.J., Zhang, M., and Spink, A. (2007). Patterns and transitions of query reformulation during web searching.

[18] Jeng, W. Jiang, J., and He, D. (2013). Users’ perceived difficulties and corresponding reformulation strategies in voice search. In Proc. HCIR, Vancouver, Canada

[19] Jiang, J., Jeng, W. and He, D. (2013). How do users respond to voice input errors? lexical and phonetic query reformula-tion in voice search. *Proc. SIGIR ’13*, 143–152.

[20] Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P. (2002). MATCH: An architecture for multimodal dialogue systems Proc. AC, 376–383.

[21] Jones, R. and Klinkner, K.L. (2008) . Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In Proc. CIKM.

[22] Lau, T. and Horvitz, E. (1999). Patterns of search: analyzing and modeling Web query refinement. In User Modeling ‘99, 119-128.

[23] C. Lucchese, S. Orlando, R. Perego, F. Silvestri and G. Tolomei. (2011). Identifying Task-based Sessions in Search Engine Query Logs. In Proc. WSDM 2011.

[24] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.

[25] Lawrence Philips. Hanging on the Metaphone. Computer Language, 12(7), 39-43, 1990.

[26] Shokouhi, M., Jones, R., Ozertem, U., Raghunathan, K., Diaz, F. (2014) Mobile query reformulations. In Proc. SIGIR. 1011-1014.

[27] Teevan, J., Adar, E., Jones, R., and Potts, M. (2007) Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In Proc. SIGIR.

[28] Tur, G., and De Mori, R. (2011). Spoken Language Understanding: Systems for extracting semantic information from speech. John Wiley & Sons.

[29] Wahlster, W. (2006). SmartKom: Foundations of multimodal dialogue systems. Springer.

[30] Young, S., Gasic, M., Thomson, B and Wiliams J. D. (2013). POMDP-based statistical spoken dialog systems: A review. Proc. IEEE, 101(5), 1160-1179.