What Do Users Say to Their TVs? An Analysis of Voice Queries to an Entertainment System

ABSTRACT

A recently-introduced product of X, a large cable company in the United States, is a "voice remote" that accepts spoken queries from users. We present an analysis of a large query log from this service, attempting to answer the question: "What do users say to their TV?" In addition to a descriptive characterization of queries and sessions, we describe two complementary types of analyses that are necessary for query understanding. First, we propose a domain-specific intent taxonomy for characterizing user behavior: as expected, most intents revolve around watching programs—both direct navigation as well as browsing—although there is a non-trivial fraction of non-viewing intents as well. Second, we propose a domain-specific tagging scheme for labeling query tokens, that when combined with intent and program prediction, offers a multi-faceted approach to understanding users' voice queries.

1 INTRODUCTION

The increasing ubiquity of intelligent agents and speech-enabled interfaces has led to a proliferation of in-home gadgets designed to address users' needs—Amazon's Echo and Google Home are two well-known examples. Recently, researchers have begun to examine voice-based interactions with entertainment systems (i.e., TVs). Rao et al. [8] introduced the problem of voice navigational queries, where users specify the program they wish to watch, e.g., "Star Trek: Discovery", and the entertainment system switches to the correct channel directly—which is more convenient than scrolling through channel guides or awkwardly trying to type in the name of the show on the remote controller. This is accomplished with hierarchical recurrent neural networks to capture session context, to recover from ASR errors, and to disambiguate short queries.

Rao et al. [8] provided a detailed look at an important but relatively narrow problem. However, in reality users can have diverse intents when talking to their TVs (i.e., tune channel, watch event, check weather, etc.), and program navigation is only a subcategory of such queries. What is missing in the literature, is a broader understanding of voice interactions between users and their entertainment systems. Literally, what are users saying to their TVs? What is the distribution of queries, query lengths, and sessions? What are the intents and user needs beyond navigational queries? What methods do we need in order to properly understand users? This paper provides a look based on a large query log of a major U.S. cable company comprising 81.M voice queries, with the aim of answering the above questions. To our knowledge, this work represents the first published analysis of such data.

The contribution of our work is twofold: First, we provide a descriptive characterization of users' voice queries along a number of standard measures such as query frequency, query length, and session length, etc. Second, we provide a methodological contribution by providing a framework for how users' voice queries in this domain should be analyzed. In particular, we propose a taxonomy of user intents and explain the need for fine-grained domain-specific query tagging. And finally, we explain how intent classification, program prediction, and query tagging present a complementary and multi-faceted approach to understanding users' voice queries.

2 BACKGROUND AND RELATED WORK

The context of our work is voice search on the X entertainment platform of BLINDED, one of the largest cable companies in the United States with approximately 22 million subscribers in 40 states. X is a software package distributed as part of X's cable box, which has been deployed to 17 million customers since around 2015. X can be controlled via the "voice remote", which is a remote controller that has an integrated microphone to receive voice queries from the customers. As expected, most queries revolve around a user's desire to watch TV, but the system has diverse capabilities beyond automatically switching channels using voice. As we shall see, there are a non-trivial fraction of non-viewing intents as well.

There is a rich body of work on voice search [1–3, 10, 11], particularly in the context of mobile devices. However, to our knowledge we are the first to analyze a large log of voice queries directed at an entertainment system. This is obviously a different context from previous studies—in our case, viewers are likely to be sitting in front of a television. To compare and contrast viewers' actual utterances, we can turn to previously-published work that studied the characteristics of voice search logs, especially in comparison to text search data [4, 9, 12]. For example, Schalkwyk et al. [9] reported statistics of queries from Google Voice's search logs and found short queries (particular 1-word and 2-word queries) to be prevalent. Interestingly, Guy[4] reported that voice queries tended to be longer than text queries, based on query logs from the Yahoo! mobile search application. We provide descriptive statistics from our query logs as a point of comparison.

Another obvious difference between the TV and other mobile devices is the display and input modality. The resolution of most TVs and their placement (i.e., distance) relative to viewers is not conducive to displaying web pages, so backing off to a generic web search for a voice query is not a desirable action. This stands in contrast to mobile search, where most modern websites adopt responsive layouts that render well on mobile devices. Furthermore, since most TV remotes (including ours) do not have a full QWERTY keyboard, users are hampered by the lack of an efficient input device for subsequent interactions with generic web pages.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WOODSTOCK'97, July 1997, El Paso, Texas USA

© 2016 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

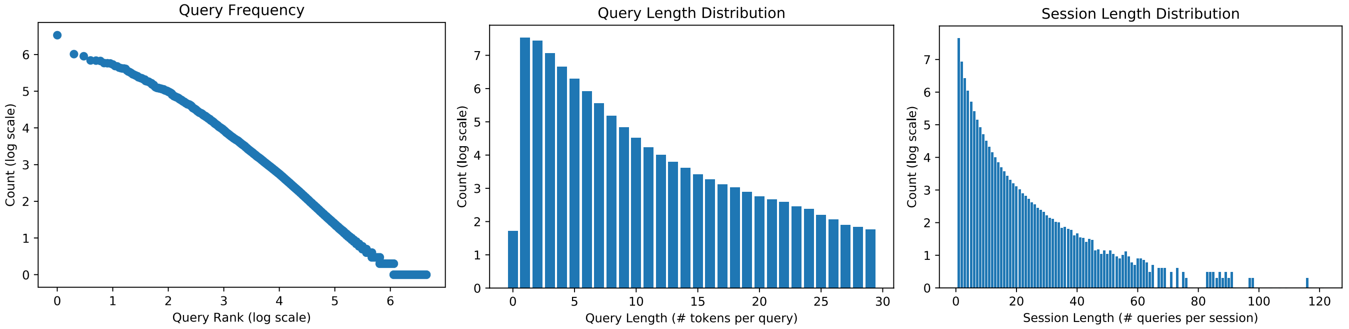

Figure 1: Distributions of voice utterances: frequency calculated with respect to raw query, query length, and session length.

Yet another difference between TVs and mobile devices is that the latter is a personal device, while the former is often shared among different members in a household. This makes personalization a challenge, since we currently have no easy way to know how is watching: for example, recommending crime dramas that may be violent would likely be inappropriate for kids in the family, but that may be perfectly relevant for the parents. Combinations of different viewers further complicate the problem.

3 LOG ANALYSIS

We present an analysis of log data collected from the X platform. This particular dataset, gathered during the week of Feb. 22 to 28, 2017, 81.4M voice queries were received from 8.1M unique users.

A few caveats are necessary to provide context: the system receives as input the one-best result of a black-box ASR system, which is a text string. We do not have access to the acoustic signal, a transcription lattice, or any additional information. In the case of X, for a variety of reasons, the ASR is outsourced to a third-party. The ASR system is specifically tuned to our domain; however, there are millions of program titles, hundreds of thousands of person names, and tens of thousands of sports team names (all of which overlap with each other at the word level) that need to be recognized, especially because television content is often very localized (e.g., viewer wants to watch local sporting event with “augsburg auggies”). This makes it non-trivial to tune ASR towards our domain.

Another challenge is the diversity of the user base in terms of age, ethnicity, etc. For example, we have observed that many ASR errors are coming from kids wanting to access their favorite cartoon. Liao et al. summarize the challenges of such ASR with children [6].

Finally, it is important to recognize that this analysis represents a (recent) snapshot in time. The model deployed today has been improved, and there is always a co-evolution of system capabilities and user queries.

3.1 Query and Session Lengths

Out of the 81.4M voice queries, there were 4.46M unique queries, indicating that despite repetition (e.g., “CNN” is a very popular query), there is a challenging amount of linguistic diversity in the data. A query has 1.96 tokens and 9.70 characters on average, and the number of unique tokens is 199K. Around 7.4 percent of tokens are out of vocabulary (OOV) words that cannot be found in the vocabulary of google word2vec [7]—13.8 percent of queries have OOV words. Most OOV words are due to ASR errors (e.g., a Canadian cartoon called “Caillou” was recognized as “cacio”).

Figure 1 presents three views of the logs. The leftmost plot shows the frequency of each unique query received within that week on a logarithmic scale (base 10). The x axis represents the rank of each query when sorted from most commonly occurring queries (“Netflix”, “CNN”, “Fox News”, “ABC”, “Free movies” are the top 5 frequent queries uttered by viewers hundreds of thousands of times per week) to the rarest ones (over 33M queries appeared only once in the logs). This is also presented in logarithmic scale to fit the entire distribution. The linearly decreasing trend on a log-log scale points to a Zipfian distribution, which is how we expect natural-language queries to be distributed when drawn from real-world applications. While the classification of query intents is explored in Section 3.2, we should note that in addition to channel names and favorite apps, some of the top queries are intended for browsing the catalog—“free movies”, “on demand”, and “movies” are among the top 20 in terms of frequency.

The center plot shows the number of tokens in a query (x axis) against its frequency (y axis). After removing punctuations and normalizing text, around 42% of incoming queries consist of a single token, many of which are single-word channel tunes. Zero-token queries consist solely of punctuations (likely ASR errors). Some of the longer queries can be quite specific movie descriptions (e.g., “go on the movie when the kids are on the bold and 22 of them got stranded on island”), or just an excited kid repeating the same query over and over (e.g., “the amazing world of gumball” repeated four times). Movie quotes and lyrics also tend to be longer in length.

Finally, the rightmost plot in Figure 1 analyzes the number of queries in a “voice session”, which we define as a sequence of consecutive voice queries with at most 45 seconds in between. The slope of decrease is much sharper in this case, compared to the center plot—in fact, more than 77% of the sessions contain only a single query. However, a considerable amount of very long sessions exists, sometimes up to hundred or more consecutive queries. Some of these tend to be exploratory, where the viewer uses voice to navigate the catalog around a central theme. Exploring the cast of a movie or a series of similarly-themed movies are two such examples. Others are more mechanic—for example, there are viewers that “zap” through channels by uttering channel names or numbers one by one (e.g., “channel 22”, “channel 20”, “cnn”, etc.). There are also few cases where the viewer appears to be having fun with the remote by saying random things.

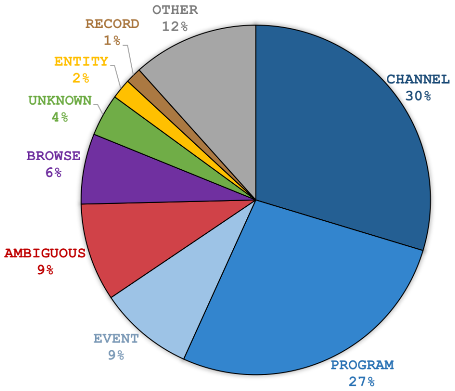

Figure 2: Intent distribution from our query logs.

3.2 Intent Classification

In this section, we introduce a taxonomy for user intents to describe different types of queries. Intent understanding can be crucial in our scenario and other home intelligence applications, since each type of intent represents different user information needs and requires a intent-specific component to handle that type of queries appropriately (i.e., watch TV vs. checking the weather). This analysis depends on the output of the production system that was deployed at a particular point in time. The system was based on a combination of hand-crafted rules and machine-learned models to detect user intents—we would characterize the accuracy as "reasonable", but certainly not perfect. Although system output error is a confound, we do not believe it would substantively alter our findings.

The distribution of these intents in our sample query log is shown in Figure 2. Not surprisingly, user queries to an entertainment system revolve around a desire to watch something. At a high level, we break this intent down into whether the user is looking for a specific program (View) or not (Browse). In our logs, the View intent comprises 66% of all queries, and can be further broken down into the following three categories:

- View Channel (30%): the viewer wishes to watch a specific channel such as "HBO" or "ESPN". These voice queries obviate the need for the viewer to remember specific channel numbers.
- View Program (27%): the viewer wishes to watch a specific program by name. This could be a series (e.g., "Game of Thrones"), a specific movie ("Back to the Future"), a comedy act, etc.
- View Event (9%): the viewer wishes to watch the broadcast of an event, such as the "Super Bowl" or the "Oscars". These events are almost always manually curated.

Browse intent represents 6% of queries, where users do not have a specific program in mind. An example might be "show me free kids movies", or "HD movies with Julia Roberts"—the viewer has some idea of the desired program but is expecting suggestions from the system. Any query that involves filtering the program catalog (no matter how broad or specific) is identified with this intent.

Beyond View and Browse, our taxonomy includes three other less common categories:

- Entity (2%): the viewer wishes to examine a particular entity profile (e.g., of an actor such as Tom Hanks). This profile includes their picture, bio, filmography, etc.
- Record (1%): the viewer is accessing DVR functions.
- Other (12%): there is a long tail of infrequent intents (a few dozen) that we lump into one category for simplicity. These include everything from toggle closed captioning, accessing the home security system, debugging wifi connections, and engaging external apps.

Finally there are two categories that are specifically artifacts of the current production system:

- Ambiguous (9%): the system identified two or more possible intents and prompts the user with a "did you mean..." dialog.
- Unknown (4%): the system was not able to identify an intent, either due to algorithmic limitations or genuine cases in which no clear intent was expressed.

The View intent is analogous to known-item retrieval in the document retrieval context and captures what Rao et al. [8] call navigational voice queries. In their formulation, these queries can be treated as multi-way classification against the entire program catalog. While these instances do dominate our query logs, we explain below why this approach alone is insufficient.

3.3 Query Tagging

There are at least two reasons why multi-way classification to predict the user's intended program falls short: First, for intents other than View, program prediction obviously make no sense. Second, even for View intents, a classification-based formulation has difficulty handling tail programs. There are typically tens of thousands of programs that are accessible to viewers at any time, especially including on-demand titles. For programs that are not frequently watched, there is insufficient training data for programs that are rarely watched. As a specific example, the model of Rao et al. [8] handles only 471 programs (based on thresholding of the training data). It would be desirable to give users voice access to the entire catalog.

To address these issues, we employ query tagging, which works in conjunction with intent classification to provide a fine-grained analysis of users' queries. Here, the problem is formulated as a sequence labeling task, where we assign a tag to each token. The tag set is as follows:

- Person: a person named entity.
- Title: the title of a program.
- Team: a sports team or sports-related term (e.g., "NFL").
- Cost: terms related to cost (e.g., "free").
- Format: terms related to format (e.g., "HD", "4K").
- Asset: e.g., "movie", "series", "music video", etc.
- Genre: e.g., "drama", "action", "comedy", etc.
- Context: a catch-all for all other terms.

For example, from the query "Watch Tom Hanks movies in HD", we extract a sequence of tags:

Context Person Person Asset Context Format

Similar to the intent detection, the current system takes advantage of a combination of handcrafted patterns and machine learning models to parse the query into the logical form (Genre="Tom Hanks" ∧ Asset="Movie" ∧ HD=`true`), which is then used to filter the program catalog to come up with a list of suggestions.

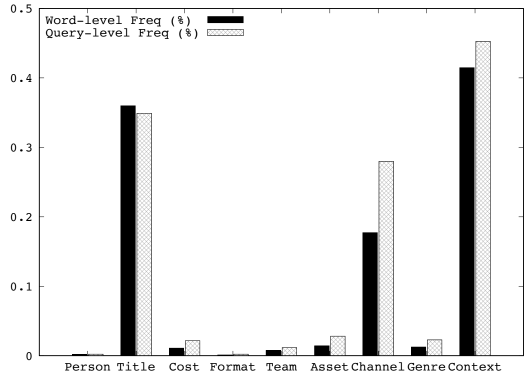

Figure 3: Distribution of query tags.

In Figure 3, the solid dark bar shows the distribution of tags over all tokens observed in our dataset, whereas the light patterned line shows percentage of queries in which each tag exists. Based on this, about 58% of tokens are part of either a named entity or modifier (not Context). Only 29% of queries are entirely made up of context tokens (i.e., no entities or modifiers were extracted). In this entity-heavy dataset, title and channel mentions alone constitute over half of all tokens. Even though some of the tag types are less frequent than others (e.g., only 1% of tokens are tagged as Genre), a good user experience requires that a wide range of capabilities (e.g., genre-based movie browsing) work reliably.

Intent classification, program prediction, and query tagging work together in a complementary way. In cases where the decision overlaps—for example, the system detects View Channel intent, which is confirmed by the tagging and program prediction—multiple sources of evidence reinforces the confidence in the decision. In cases where program prediction fails—for example, rarely-watched programs—tagged tokens in the user's query can serve as keywords for search our program catalog. Finally, for an intent such as Browse, the various modifiers from tagging (e.g., Format, Cost, etc.) play an important role to address the user's information needs. This represents a case where intent prediction and query tagging need to work together to understand a user's query.

3.4 Beyond Navigational Queries

Finally, we present a preliminary linguistic analysis of our query logs to provide a glimpse into what users of a voice-enabled entertainment system are looking for beyond navigational queries.

In order to rank queries based on some "naturalness" measure, we trained a language model using the Hansard parliament speech corpus (0.76M sentences) and the IMDB movie review dataset (1.22M sentences). As a filtering step, we removed all queries that exactly matched a title in our catalog, as well as any query with five tokens or less. What was left was a set of 2.9M queries (1.1M unique), which were then scored by the language model and sorted by the LM score plus the log-frequency. This yields a ranked list of frequently occurring natural utterances directed at the voice remote.

Analyzing the resulting set, we noticed a wide range of intents. In fact, the percentage of Unknown queries were 50% higher in this subset of the logs, pointing to an increased level of complexity. Also, the percentage of Browse queries was much higher (15% vs. 6%), emphasizing the need for a tagging-based approach (as presented in Section 3.3) to properly understand this long tail of queries.

Queries in the top of the list ranged from movie quotes and music lyrics (e.g., "All i want to say is that they don't really care about us") to very specific requests (e.g., "Return to the movie that I did not finish last night"). On the other hand, there were also open-ended questions (e.g., "Do you have a movie about the Vietnam War?") as well as factual questions (e.g., "Who is being nominated for best picture in the academy awards?").

To gain a little more insight into the syntactic structure, we ran a dependency parser [5] on all 1.1M unique queries. The most common root word is "show" with part-of-speech "verb" (`show/VB`), comprising 12% of all queries in this subset. In fact, root words of the verb form (`VB`, `VBP`, `VBZ`, etc.) made up over half of all queries. The remaining queries have a root with part-of-speech noun (40%), adjective (2%), preposition (1%), determiner/pronoun (negligible). The most frequently observed noun root was `movies/NNS`—for adjective and prepositions, `free/JJ` and `on/IN` topped the list.

4 CONCLUSIONS

Work on building models that understand voice queries in an entertainment context is still very much in its infancy, as being able to talk to a TV remains a novel feature for most consumers (in contrast to mobile devices, where consumers already have an expectation of voice-based interactions with intelligent agents). Although there is much to learn and low hanging fruit in applying well-known techniques from web search, the very different context necessitates new techniques. In this paper, we present three small steps in that direction: a descriptive characterization of users' voice queries on TV, a domain-specific intent taxonomy and query tagging scheme. Nevertheless, there remains many unique challenges left unexplored, and we believe the entertainment domain will prove to be a fertile ground for future research.

REFERENCES

[1] A. Acero, N. Bernstein, R. Chambers, Yun-Cheng Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig. 2008. Live Search for Mobile: Web Services by Voice on the Cellphone. In *ICASSP.*

[2] C. Chelba and J. Schalkwyk. 2013. Empirical Exploration of Language Modeling for the google.com Query Stream as Applied to Mobile Voice Search. In *Mobile Speech and Advanced Natural Language Solutions.*

[3] J. Feng and S. Bangalore. 2009. Effects of Word Confusion Networks on Voice Search. In *EACL.*

[4] I. Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *SIGIR.* 35–44.

[5] L. Kong, C. Alberti, D. Andor, I. Bogatyy, and D. Weiss. 2017. DRAGNN: A Transition-based Framework for Dynamically Connected Neural Networks. (03 2017). arXiv:1703.04474 https://arxiv.org/abs/1703.04474

[6] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q. Jiang, T. Sainath, A. Senior, F. Beaufays, and M. Bacchiani. 2015. Large vocabulary automatic speech recognition for children. In *Interspeech.*

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS.*

[8] J. Rao, F. Ture, H. He, O. Jojic, and J. Lin. 2017. Talking to Your TV: Context-Aware Voice Search with Hierarchical Recurrent Neural Networks. In *CIKM.* 557–566.

[9] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. 2010. "Your Word is My Command": Google Search by Voice: A Case Study. In *Advances in Speech Recognition.*

[10] J. Shan, G. Wu, Z. Hu, X. Tang, M. Jansche, and P. Moreno. 2010. Search by Voice in Mandarin Chinese. In *INTERSPEECH.*

[11] Y. Wang, D. Yu, Y. Ju, and A. Acero. 2008. An Introduction to Voice Search. *IEEE Signal Processing Magazine* 25, 3, 29–38.

[12] J. Yi and F. Maghoul. 2011. Mobile Search Pattern Evolution: The Trend and the Impact of Voice Queries. In *WWW.*