What Do Viewers Say to Their TVs? An Analysis of Voice Queries to Entertainment Systems

Jinfeng Rao,[1,2] Ferhan Ture,[1] and Jimmy Lin[3]
[1] Comcast Applied AI Research Lab
[2] Department of Computer Science, University of Maryland
[3] David R. Cheriton School of Computer Science, University of Waterloo
jinfeng@cs.umd.edu,ferhan_ture@comcast.com,jimmylin@uwaterloo.ca

ABSTRACT

A recently-introduced product of Comcast, a large cable company in the United States, is a "voice remote" that accepts spoken queries from viewers. We present an analysis of a large query log from this service to answer the question: "What do viewers say to their TVs?" In addition to a descriptive characterization of queries and sessions, we describe two complementary types of analyses to support query understanding. First, we propose a domain-specific intent taxonomy to characterize viewer behavior: as expected, most intents revolve around watching programs—both direct navigation as well as browsing—but 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, provides a multi-faceted approach to understand voice queries directed at entertainment systems.

ACM Reference Format:
Jinfeng Rao, Ferhan Ture, and Jimmy Lin. 2018. What Do Viewers Say to Their TVs? An Analysis of Voice Queries to Entertainment Systems. In *SIGIR '18: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8-12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210140

1 INTRODUCTION

The increasing ubiquity of intelligent agents and speech-enabled interfaces has led to a proliferation of in-home gadgets designed to address various user needs—Amazon's Echo and Google Home are two well-known examples. It makes sense that users would also like to engage in voice-based interactions with their entertainment systems (i.e., TVs). This has been the focus of our work: as a first step, we recently introduced the problem of voice navigational queries [8], where viewers specify the program they wish to watch, e.g., "Star Trek: Discovery", and the entertainment system switches to the correct channel. Such interactions are more convenient than scrolling through channel guides or awkwardly trying to type in the name of the show on the remote controller.

To tackle this challenge, we described the use of hierarchical recurrent neural networks to capture session context, recover from ASR errors, and disambiguate short queries [8]. While important, this is a relatively narrow problem. In reality, viewers have diverse intents when talking to their TVs (i.e., change channel, record show, check weather, etc.) and program navigation is only one category of such queries. What is missing in the literature is a broader understanding of voice interactions between users and entertainment systems. Literally, what are viewers saying to their TVs? What is the distribution of queries, query lengths, and sessions? What are the intents expressed by viewers beyond navigational queries? What do we need in order to properly understand user needs? This paper provides a look based on a large query log from Comcast, with the aim of answering the above questions. To our knowledge, this work represents the first published analysis of such data.

Our work makes two contributions: First, we provide a descriptive characterization of viewers' voice queries along a number of standard measures such as query frequency, query length, and session length. Second, we provide a methodological contribution by presenting a framework for how entertainment voice queries should be analyzed. In particular, we propose a taxonomy of user intents and explain the need for fine-grained domain-specific query tagging. Finally, we discuss how intent classification, program prediction, and query tagging form a complementary and multi-faceted approach to understand entertainment voice queries.

2 BACKGROUND AND RELATED WORK

The context of our work is voice search on the X1 entertainment platform by Comcast, one of the largest cable companies in the United States with approximately 22 million subscribers in 40 states. X1 is a software package distributed as part of the company's cable box, which has been deployed to 17 million customers since around 2015. The platform can be controlled via the "voice remote", which is a remote controller with an integrated microphone to receive spoken queries from customers. As expected, most queries revolve around watching TV, but the system has diverse capabilities beyond switching channels by voice. As we shall see, there is 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 dataset of voice queries directed at entertainment systems. This is obviously a different context compared to mobile devices—in our case, customers are likely to be sitting in front of a television. As a starting point to understanding these queries, we can turn to previous studies analyzing voice search logs, especially in comparison to text search [4, 9, 12]. For example, Schalkwyk et al. [9] reported statistics of queries from

SIGIR'18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210140

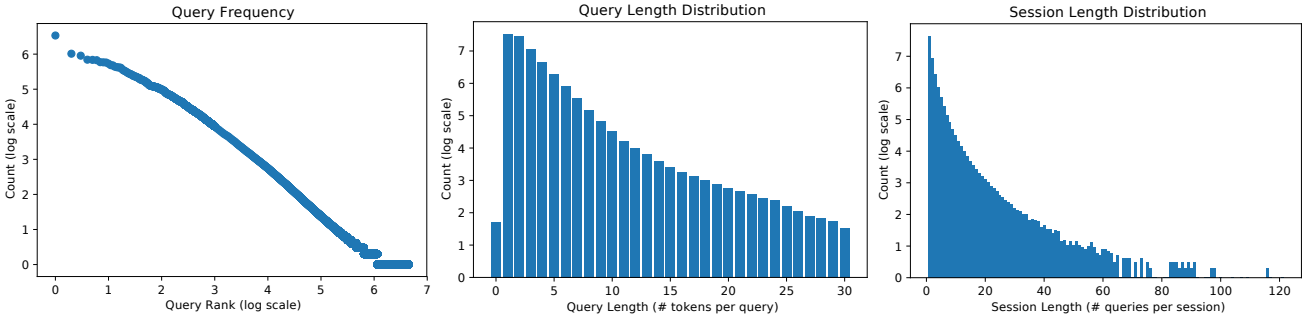

Figure 1: Characteristics of voice queries directed at entertainment systems: distribution of query frequency (left), query length (middle), and session length (right).

Google Voice's search logs and found short queries (particular one- and two-word queries) to be prevalent. Interestingly, Guy [4] reported that voice queries tend to be longer than text queries, based on Yahoo! mobile query logs. We provide descriptive statistics from our query logs as a point of comparison.

An obvious difference between TVs 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, viewers are hampered by the lack of an efficient input device for subsequent interactions with web pages.

Yet another difference between TVs and mobile devices is that the latter is highly personal, whereas the former is typically shared among different members in a household. This makes personalization a challenge, since we currently have no easy way to know who is watching. For example, recommending (potentially violent) crime dramas would likely be inappropriate for kids in the family, but that may be exactly what the parents want. Combinations of different viewers further complicate the problem.

3 LOG ANALYSIS

We present an analysis of log data collected from the X1 platform during the week of Feb. 22 to 28, 2017. Our dataset contains 81.4M voice queries from 8.1M unique devices.

A few caveats are necessary to provide context: our system receives as input the one-best result of a black-box third-party ASR system, which is a text string. We do not have access to transcription lattices or n-best lists. Although the ASR system is specifically tuned to our domain, it needs to recognize millions of program titles, hundreds of thousands of person names, and tens of thousands of sports teams, all of which overlap with each other. Television content is often very localized, e.g., a viewer wants to watch local sporting event with the "Augsburg Auggies", making domain adaptation difficult.

Another challenge is the diversity of customers in terms of age, ethnicity, etc. For example, we have observed that many ASR errors come from kids wanting to access their favorite cartoon; see Liao et al. [6] for a summary of ASR challenges with children.

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 customer queries.

3.1 Query and Session Lengths

Out of the 81.4M voice queries, there are 4.46M unique queries, indicating that despite the presence of frequent head queries (e.g., "CNN"), there is plenty 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 (constrained by the ASR system). Around 7.4% of tokens are out of vocabulary (OOV) with respect to the Google News corpus used to train word2vec [7]; 13.8% queries have OOV words. Most OOV words are due to a mismatch between the vocabulary of the ASR system and our text processing tools.

Figure 1 presents three views of our dataset. The left panel shows a standard log–log (base 10) plot of query frequencies. The top five most frequent queries are "Netflix", "CNN", "Fox News", "ABC", and "free movies", uttered by viewers hundreds of thousands of times per week. In the tail, we observe 3.3M unique queries. Unsurprisingly, the distribution is Zipfian. We examine query intents in more detail in Section 3.2, but note here that in addition to channel names and favorite apps, some of the most frequent queries are intended for browsing the catalog, where the viewer does not have a specific program in mind; "free movies", "on demand", and "movies" are among the top 20 in terms of frequency.

The center panel shows the distribution of query lengths in terms of the number of tokens; for clarity, we only show queries with lengths up to 30 tokens. After removing punctuations and normalizing text, around 42% of incoming queries consist of a single token, many of which are single-word channel names. Zero-length queries, comprised solely of punctuations, are mostly 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). We also recognize movie quotes and lyrics in our dataset, which tend to be longer in length.

Finally, the right panel in Figure 1 shows the number of queries in a "voice session", which we define as a sequence of consecutive queries with a maximum gap less than 45 seconds between queries. More than 77% of the sessions contain only a single query. However,

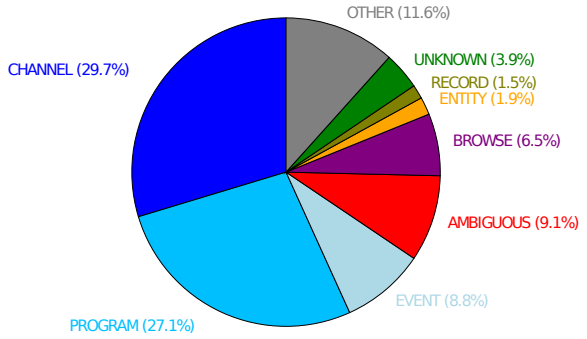

Figure 2: Intent distribution from our query logs.

a considerable number of very long sessions exist, sometimes up to a hundred or more 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 mechanical—for example, there are viewers who "zap" through channels by uttering channel names or numbers one by one (e.g., "channel 22", "channel 20", "CNN", etc.). There are also cases where the viewer appears to be having fun with the remote by saying random things.

3.2 Intent Classification

In this section, we introduce a taxonomy for viewer intents to categorize different types of queries. Note that our analysis is based 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 viewer intents, over a taxonomy that has organically evolved over time. We would characterize the accuracy as "reasonable", but certainly not perfect. Although system output error is a confound, we do not believe errors substantively alter our findings.

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

- View Channel (29.7%): 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 (27.1%): 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 (8.8%): 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.

The Browse intent, where viewers do not have a specific program in mind, represents 6.5% of queries. Examples are "show me free kids movies" or "HD movies with Julia Roberts". In these cases, the viewer has some idea of the desired program but is expecting suggestions from the system. Any query that involves filtering the program catalog is identified with this intent.

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

- Entity (1.9%): the viewer wishes to examine a particular entity profile (e.g., of an actor such as Tom Hanks). This profile includes the actor's picture, bio, filmography, etc.
- Record (1.5%): the viewer is accessing DVR functions.
- Other (11.6%): there is a long tail of infrequent intents (a few dozen) that we lump together. These include everything from toggling 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 production system:

- Ambiguous (9.1%): the system identified two or more possible intents and prompts the viewer with a "did you mean..." dialog.
- Unknown (3.9%): 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 we have previously called navigational voice queries [8].

3.3 Query Tagging

In our previous work [8], query understanding is formulated as multi-way classification over a set of programs. Although queries with the View intent dominate our dataset, there are at least two reasons why such an approach falls short: First, for intents other than View, program prediction obviously makes no sense. Second, even for View intents, a classification-based formulation has difficulty handling tail programs. There are typically tens of thousands of programs accessible to viewers at any time, especially including on-demand titles. For programs that are not frequently watched, there is insufficient training data; for example, our previous model handles only less than a thousand programs. It would be desirable to give viewers 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 viewers' queries. Here, the problem is formulated as a sequence labeling task, with the following tag set:

- 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 intent detection, the current system takes advantage of handcrafted patterns as well as machine-learned models to parse the query into the logical form, e.g., (Person="Tom Hanks" ∧ Asset="Movie" ∧ Format="HD"). This is then used to filter the program catalog to provide a list of suggestions.

In Figure 3, the solid dark bars show the distribution of tags over all tokens observed in our dataset, whereas the lighter gray bars

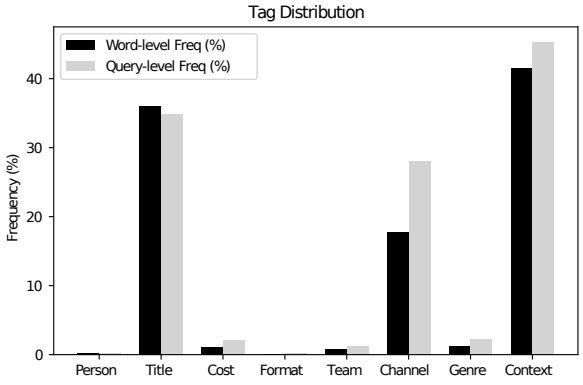

Figure 3: Distribution of query tags.

show the percentage of queries in which each tag exists. Based on this, we see that 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 occur less frequently than others (e.g., only 1% of tokens are tagged as Genre), high accuracy for all tags are necessary to produce a good user experience. For example, genre-based movie browsing requires reliably identifying Genre tags.

Intent classification, program prediction, and query tagging work together in a complementary fashion. 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 reinforce the confidence in the decision. In cases where program prediction fails—for example, rarely-watched programs—tagged tokens in the query can serve as keywords for searching the program catalog. Finally, for an intent such as Browse, the various modifiers from tagging (e.g., Format, Cost, etc.) play an important role to understand a viewer's query. In this example, intent prediction and query tagging need to work together to generate an appropriate response.

3.4 Beyond Navigational Queries

Finally, we present a preliminary linguistic analysis of our query logs to provide a glimpse into the diversity of viewer queries posed to entertainment systems. In order to score 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 matched a title in our catalog exactly as well as any query with five tokens or less. This yielded 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 of the frequency of occurrence. The result is a ranked list of frequently-occurring "natural" utterances directed at the voice remote.

Analyzing the results, we observe a wide range of intents. In fact, the percentage of Unknown queries is 50% higher in this subset of the logs, pointing to an increased level of complexity. The percentage of Browse queries is also much higher (15% vs. 6%), which affirms the need for a tagging-based approach (as presented in Section 3.3) to properly understand complex queries.

Queries ranked highly in our "naturalness" measure 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 of the queries, we ran a dependency parser [5] on all 1.1M unique queries in this subset. The most common root word was "show" with part-of-speech verb (`show/VB`), comprising 12% of all queries. In fact, root words of verb forms (`VB`, `VBP`, `VBZ`, etc.) comprised half of all queries. The remaining queries had a root with the part-of-speech noun (40%), adjective (2%), preposition (1%), and determiner/pronoun (negligible). The most frequently observed noun root was `movies/NNS`; for adjective and prepositions, `free/JJ` and `on/IN` topped the list, respectively.

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 an expectation of voice-based interactions with intelligent agents on mobile devices). Although there is much to learn and low hanging fruit in applying mature 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 viewers' voice queries on TVs, a domain-specific intent taxonomy, and a query tagging scheme. Nevertheless, there remain many unexplored challenges, and we believe the entertainment domain will prove to be a fertile ground for future research.

REFERENCES

[1] A. Acero, N. Bernstein, R. Chambers, Y. C. Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig. 2008. Live Search for Mobile: Web Services by Voice on the Cellphone. *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. *Mobile Speech and Advanced Natural Language Solutions*.
[3] J. Feng and S. Bangalore. 2009. Effects of Word Confusion Networks on Voice Search. *EACL*.
[4] I. Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. *SIGIR*.
[5] L. Kong, C. Alberti, D. Andor, I. Bogatyy, and D. Weiss. 2017. DRAGNN: A Transition-based Framework for Dynamically Connected Neural Networks. arXiv: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. *Interspeech*.
[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *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. *CIKM*.
[9] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Garrett, and B. Strope. 2010. "Your Word is My Command": Google Search by Voice: A Case Study. *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. *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. *WWW*.