Challenges and Opportunities in Understanding Spoken Queries Directed at Modern Entertainment Platforms

Ferhan Ture,[1] Jinfeng Rao,[1] Raphael Tang,[1,2] and Jimmy Lin[2*]
[1] Comcast Applied AI Research Lab
[2] David R. Cheriton School of Computer Science, University of Waterloo

1 INTRODUCTION

Modern in-home entertainment platforms—representing the evolution of the humble television of yesteryear—are packed with features and content: they offer a dizzying array of programs spanning hundreds of channels as well as a catalog of on-demand programs offering tens of thousands of options. Furthermore, the entertainment platform may serve as an in-home hub, providing capabilities ranging from playing music to controlling the home security system. At a high level, our goal is to provide natural speech-based access to these myriad features as an alternative to physical button entry on a remote control.

The Comcast Xfinity X1 entertainment platform supports speech-based interactions via a voice-enabled remote controller and set-top box, which displays feedback on the television. Comcast has delivered more than 20 million voice-enabled remotes to customers across the United States, processing more than 9 billion voice commands in 2018. The operation is quite simple: the viewer presses a microphone button, issues a voice query, and then releases the button. The utterance is transcribed by an automatic speech recognition (ASR) system, fed to a number of natural language understanding modules, and the TV responds appropriately. For example, the viewer might say "watch Inception on demand" and the TV will tune to the proper listing. The system also supports browsing queries such as "show me free kids movies" or "movies with Scarlett Johansson", as well as many other intents.

This abstract provides an overview of recent work at the Comcast Applied AI Research Lab (along with collaborators) to highlight challenges and opportunities in building intelligent agents for the entertainment vertical.

2 HOW IS THIS DIFFERENT?

The many previous studies on voice queries, mostly in the context of web search on smartphones, have shown that speech input is more than just "slapping ASR" in front of text search. Voice queries differ in length and other descriptive characteristics [3, 8], and appear more natural than text queries [3]. They exhibit different patterns of query reformulations [1, 4, 9], and are affected by issues not present in text-based input, such as speaker variability and environment noise [12]. We believe that voice interactions with entertainment systems differ in key ways from personal assistants on smartphones, smart speakers, and other voice-enabled devices, and that this vertical deserves dedicated attention.

In our case, users are typically sitting in front of their TVs, expressing an entertainment intent (or one related to the capabilities of the system, which is a closed universe). While there are open-domain aspects (for example, viewers asking trivia questions), user needs are more circumscribed than in general web search.

One obvious difference between TVs and mobile devices is the display and input modality. In voice search on smartphones, systems often back off to generic web search when pre-defined "search cards" cannot address the user's need [3]. This is not a desirable strategy in our context because TVs are typically placed at a distance that makes webpages difficult to read. Furthermore, follow-up interactions with generic web pages are difficult since the input options are limited; for example, clicking links is awkward. Other than a few limited possibilities (e.g., directional arrows and keypad), voice is the only option for interactions, and certainly the only practical modality for freeform input.

Another difference is that while mobile devices are usually personal, TVs are usually communal, shared among a household. Our analyses confirm that watching TV is often a shared social experience [10], for example, with multiple people engaging with the voice remote. Background noises such as laughter or multiple overlapping utterances are not uncommon. Not only does this present challenges to speech recognition and intent understanding, it also makes personalization quite difficult, since it is non-trivial to determine *who* is expressing a particular need and *for whom* the system is preparing a result or making a recommendation.

3 RECENT WORK

We provide a number of short "vignettes" that highlight our recent work in understanding spoken queries directed at the X1 platform:

A taxonomy of intents. In the web context, Broder's taxonomy [2] has been highly influential in understanding search queries, and the large body of literature studying the behavior of web searchers can trace its origins back to this work.

Along these same lines, as an initial step, we have proposed a domain-specific intent taxonomy, based on analysis of a large voice query log, to characterize spoken queries that are directed at our entertainment platform [7]. It is perhaps not surprising that most intents revolve around a desire to watch various entertainment programs. For approximately two thirds of queries, customers have a specific program they wish to watch, divided between channels (e.g., "CNN"), programs (e.g., "Game of Thrones"), and to a lesser extent special events (e.g., "Oscars"). Around another six percent of queries express an intent to view a program; these are vague and

*J. Rao conducted research at Comcast while a Ph.D. student at U. Maryland.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331433

represent browsing behavior—i.e., the customer is uncertain about what to watch. Examples include "free on-demand action movies" or "HD kids shows".

Beyond specific and browsing-oriented viewing intents, we observe a long tail of intents totaling around one hundred identifiable categories in our taxonomy. These include accessing DVR (recording) functions, factoid-like questions about movies and actors, and occasionally people talking to their TVs for their own amusement.

From clickthroughs to "watchthroughs". Clickthroughs in the web context represent an important behavior that underlie much of modern approaches to modeling and understanding the behavior of web users. In the entertainment domain, we can analogize the notion of "watchthroughs": users express an intent (which can be classified in terms of the taxonomy above), interact with the system for a number of rounds, and then watch a program for an extended period of time. Taken together, these interactions can be characterized as a voice search session, and similar to how web search sessions can be mined for training data based on clickthroughs, we can automatically build large implicitly-labeled datasets from viewers' queries and the programs they ultimately watch [5]. As in the web context, watchthroughs can be used to characterize system successes and failures in an unsupervised manner [10].

Neural models for query intent understanding. Prior to our recent efforts that take advantage of deep learning, the production system comprised a cascade of simple NLP modules (based on pattern matching and some machine-learned components). This is adequate for high-frequency "head" queries; for example, pattern matching suffices for queries like "CNN" or "NBC".

Our first attempt at applying modern data-driven techniques, described in Rao et al. [5], focused on directly identifying the program a customer intends to watch. We call these *navigational voice queries*, much like their counterparts in general web search. For these queries, session context can be exploited for disambiguation and to recover from ASR errors. For example, a query "game of throw" can either refer to the television series "Game of Thrones" (because of a transcription error) or a TV game called "Fish Throw Game". However, if the user just uttered "HBO series" a moment ago, then it's more likely that the user is looking for the former since we know the show is playing on HBO.

These insights are operationalized using Navigational Hierarchical Recurrent Neural Networks (N-HRNN), which was deployed into production in January 2018 [6]. This model runs as part of a cascade after a number of simpler NLP modules, to handle the "tough queries". Prior to deploying the N-HRNN, the production system was unable to produce any response for approximately 8% of all queries—in this case, the customer sees a special "cannot handle this query" message. After deployment, the coverage of the entire end-to-end system increased from 92% to 98%. In other words, the model increased coverage and reduced the number of unhandled queries by three quarters. Manual evaluation on a sample of queries shows that for two thirds of queries handled by the N-HRNN, the customer experience improved overall.

Neural models for voice query recognition. Our most recent work attempts to exploit the fact that Comcast has end-to-end control over the entire technology stack, from speech recognition to output presentation. Most voice-enabled intelligent agents, such as Apple's Siri and the Amazon Echo, are powered by a combination of two speech technologies: lightweight keyword spotting (KWS) to detect a few pre-defined phrases (e.g., "Hey Siri") and full automatic speech recognition (ASR) to transcribe complete user utterances (the X1 platform currently uses a third party for speech transcription). We wished to explore a middle ground: techniques for voice query recognition capable of handling a couple of hundred commands. This is particularly interesting because of the Zipfian distribution of users' queries: the 200 most popular queries cover a significant portion of monthly voice traffic.

We recently introduced a novel, resource-efficient neural network for voice query recognition that is both more accurate than state-of-the-art CNNs, yet can be easily trained and deployed with limited resources [11]. On an evaluation dataset, we achieve a low false alarm rate of 1% and a query error rate of 6%; the model performs inference 8.24× faster than the current ASR system.

4 OPPORTUNITIES AND CHALLENGES

The goal of our presentation is to introduce the SIGIR community to challenges, opportunities, and recent work on the problem of speech-based interactions with entertainment platforms. This domain ties into emerging areas of interest to the community—search as conversation, multi-modal input, and interactions with intelligent agents. Although some techniques from web search can be directly applied to our queries (for example, the analogy between clickthrough and watchthrough), we believe that our domain requires an even greater emphasis on query understanding, and stands to benefit from techniques from natural language processing. Since speech is the only practical input modality, it is not only beneficial, but perhaps necessary, to optimize for the end-to-end experience, from voice input to system response. Overall, TV-based entertainment systems (if we include similar systems such as the Amazon Fire TV and Roku TV) already serve tens of millions of customers today—building speech-enabled intelligent agents in this vertical is an exciting and impactful research direction!

REFERENCES

[1] A. Awadallah, R. Kulkarni, U. Ozertem, and R. Jones. 2015. Characterizing and Predicting Voice Query Reformulation. In *CIKM*. 543–552.
[2] A. Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.
[3] I. Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *SIGIR*. 35–44.
[4] J. Jiang, W. Jeng, and D. He. 2013. How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search. In *SIGIR*. 143–152.
[5] 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.
[6] J. Rao, F. Ture, and J. Lin. 2018. Multi-Task Learning with Neural Networks for Voice Query Understanding on an Entertainment Platform. In *KDD*. 636–645.
[7] J. Rao, F. Ture, and J. Lin. 2018. What Do Viewers Say to Their TVs? An Analysis of Voice Queries to Entertainment Systems. In *SIGIR*. 1213–1216.
[8] 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. In *Advances in Speech Recognition*.
[9] M. Shokouhi, R. Jones, U. Ozertem, K. Raghunathan, and F. Diaz. 2014. Mobile Query Reformulations. In *SIGIR*. 1011–1014.
[10] R. Tang, F. Ture, and J. Lin. 2018. Yelling at Your TV: An Analysis of Speech Recognition Errors and Subsequent User Behavior on Entertainment Systems. In *SIGIR*.
[11] R. Tang, G. Yang, H. Wei, Y. Mao, F. Ture, and J. Lin. 2018. Streaming Voice Query Recognition using Causal Convolutional Recurrent Neural Networks. *arXiv:1812.07754*.
[12] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero. 2008. An Introduction to Voice Search. *IEEE Signal Processing Magazine* 25, 3, 29–38.