INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Speech-based Web Navigation for Limited Mobility Users

Vasiliy Radostev, Serge Berger, Justin Tabrizi, Pasha Kamyshev, Hisami Suzuki

AI and Research, Microsoft Corporation
{vasiliyr, sergeb, jutabriz, pakamysh, hisamis}@microsoft.com

Abstract

We present a novel approach that introduces the strengths of voice assistants into a web browser that makes the task of web navigation a lot more accessible to all users, especially under limited mobility circumstances. Voice assistants have now been widely adopted and is providing great user experience for getting simple actions done quickly or getting a quick answer to a question. On the other hand, the benefits of voice assistants have not yet penetrated to the scenarios such as web navigation, which has so far been driven by mouse, keyboard and touch-based input only. In this paper, we demonstrate our speech-based web navigation system, and show that our system improves the completion of the web navigation task on both PC and mobile phone significantly as compared with an out-of-the-box voice assistants on this task.

Index Terms: web navigation, voice assistance, limited mobility

1. Introduction

Intelligent voice assistants, such as Amazon's Alexa, Apple's Siri, Google's Assistant and Microsoft's Cortana, have now been widely adopted and is providing great user experience for primarily voice-based scenarios such as getting simple actions done quickly (e.g., "set a timer for 5 minutes"; "play jazz music") or getting a quick answer to a question (e.g., "weather tomorrow"; "what time is it in Austria"). Studies also show that these voice assistants enhance the user experience of mobility impaired users [1]. Navigation to a web page (e.g. open en.wikipedia.org or youtube.com), on the other hand, is a task that is typically conducted and completed using a web browser which assumes input by typing and clicking, and do not natively support speech input; it has therefore not previously rendered itself to the benefits of voice assistants. In this paper, we show that a web navigation task can be greatly facilitated by a browser extension that incorporates voice assistant-like capabilities: our experiments show that the task completion rate of web navigation improves significantly on both PC and mobile phones with voice assistant capabilities when the input modality via a mouse, keyboard or touch is not accessible.

2. System Architecture

Figure 1 illustrates the proposed system architecture. As is similar to traditional voice assistants, the system utilizes Automatic Speech Recognition (ASR), Text-to-Speech (TTS) and Natural Language Understanding (NLU) components; we have built these assistant components into a web browser extension, which calls the NLU component to understand the intent and slots for the incoming user query (e.g. tags the query "open global warming on Wikipedia" with intent: *web_navigation*, *slot_topic*:global warming and *slot_website*: Wikipedia) and executes the web navigation action on a browser. Figure 2 shows the screenshots of the output of the browser extension we implemented on a mobile phone: user says "Go to Wikipedia" to be directly navigated to the website (A); "global warming" the finds the relevant page (B).

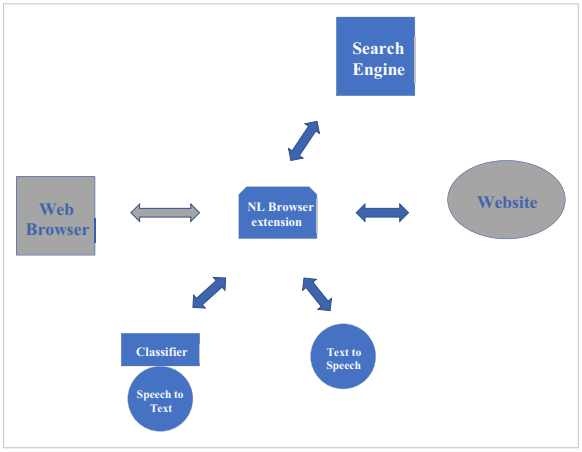

Figure 1: *System Architecture for Speech-based Web Navigation*

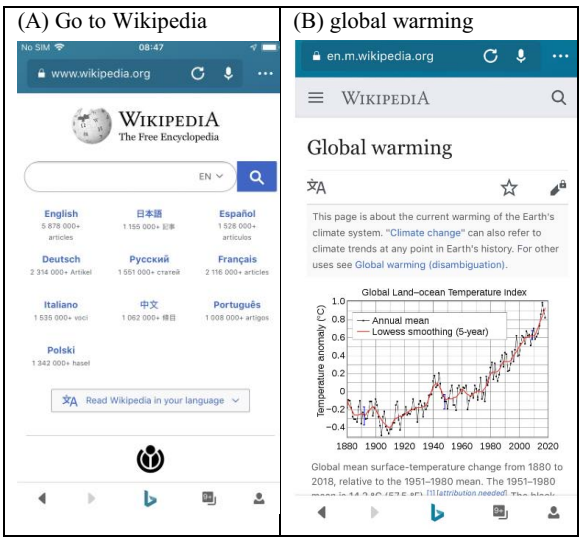

Figure 2: *Screenshots of browser (Bing Search app) extension with voice assistant capability*

Despite that the performance of some ASR systems are said to outperform the human levels based on some metrics [2], it is often the case that the users of voice assistants perceive they fail to complete a task due to speech recognition errors. For this reason, our system is designed to take n-best results

from the ASR component, and is set up to jointly optimize ASR, NLU and web navigation results for the best end-to-end accuracy. A classifier is used to decide whether to take the best ASR results or any other alternatives. We have selected only universally consistent algorithms to make sure that the classifier would scale, namely SVM and AdaBoost. A classification method is universally consistent if the risk of the classifiers it produces approaches the Bayes risk – the minimal risk – as the sample size grows. Since data returned by the engine may be sparse, we use SMOTE [3] for over-sampling and optimized hyper-parameters throughout (cf.[4]). Table 1 shows the results of an offline evaluation on the improvements in Word Error Rate (WER) Loss when experimenting on 10K labeled data by performing a 10-fold cross validation experiment with 70%/25% training and test data split. As the table shows, WER Loss improved by 2.39%. Though this classifier was not used as part of the Experiment in Section 3, we plan to use it in the near future for improving the web navigation task as well.

Table 1: *Results of the boosting classifier*

System	WER Loss	Boost
Baseline	1.42	
+Classifier	1.38	2.39%

3. Experiments

In this section we describe the experiments we ran to evaluate how the proposed system facilitates the completion of web navigation task.

3.1. Task Set Up

Our experiment was set up to measure how effectively a user can complete a web navigation task with different input modalities using different tools. The task judges were given the following instruction:

> Yesterday, you saw your friend visiting a web page *A*. You want to see the information on that page yourself today, but you can't use your hands to type. Recently, you learned there is a new feature, which allows you to search websites using voice: you can open a website *A* using speech, and follow up with a search query *B* on that website. You want to try using it today and see if you can get to the same information by voice only.

Some sample pairs of website *A* and search query *B* that the judges were given are:

A: IRS *B:* my refund

A: Amazon.com *B:* poo pourri spray

Judges can formulate speech queries freely for A and B, as in:

A: IRS homepage

B: Where is my refund

Our test set consists of 30 pairs of websites *A* and search queries *B*. The set was assembled based on a random sample of web navigation usage data from the users who provided consent to Microsoft to collect and analyze such data. Judges were given these pairs, and were asked to complete the web navigation task in 3 settings:

S1: Web browser with traditional mouse and keyboard input. We used Edge browser on PC, and Bing app on mobile phone.

S2: Voice assistant. We used Cortana on PC, and Google Assistant on mobile phone with speech input.

S3: Proposed system which combines web navigation with voice assistance.

Judges rated each web navigation task according to the following 5-scale criteria:

5 Works great, navigated to the right page

4 Got similar information, easy

3 It took me a few attempts, but I found the right information

2 It took me a few attempts, and I found similar information

1 Can't get the information using voice only

2 and 3 cover the cases where speech recognition fails in the initial attempts. We converted these ratings into scores between 100 and 0 (5 to 100, 4 to 80, 3 to 60, 2 to 40, and 1 to 0).

3.2. Results

Table 2 shows the results of our experiment. The ratings are averaged over 30 pairs across judges:

Table 2: *Results of web navigation task completion*

System	PC	Mobile
S1 (browser)	100	96
S2 (assistant)	7	22
S3 (proposed)	52	48

The results show that the web navigation task is easy to complete on a browser using mouse, keyboard or touch as input modality (S1). However the task completion rate drops significantly when the user cannot have access to mouse or keyboard input (S2) under limited mobility circumstances: the completion rate is very bad on PC (only 7%), is slightly better on mobile phone (still 22%), possibly due to the fact that voice assistants are more optimized on a mobile phone than on a PC. Our system (S3) significantly boosts the task completion rate over S2, suggesting that the web navigation can be made a lot more accessible to all users using voice assistant capabilities.

4. Discussion

The system described in this paper is integrated in Microsoft Bing Search app for mobile phones, available for download at https://apps.apple.com/us/app/bing-search/id345323231 (iOS) and https://play.google.com/store/apps/details?id=com.microsoft.bing&hl=en_US (Android); a version for PC is upcoming. We plan to continue improving the task completion rate of web navigation by improving the performance of the components of the system, as well as based on real user feedback.

5. References

[1] A. Vtyurina et al., "Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search", To appear in *Proceedings of ASSETS* 2019.

[2] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," arxiv preprint, vol.arXiv:1610.05256, 2016.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Intelligence Research*, vol. 16, pp. 321–357, 2002.

[4] I. Steinwart, "Support vector machines are universally consistent," *Journal of complexity*, vol. 18, pp. 768–791, 2002.