Life After Speech Recognition: Fuzzing Semantic Misinterpretation for Voice Assistant Applications

Yangyong Zhang, Lei Xu, Abner Mendoza, Guangliang Yang, Phakpoom Chinprutthiwong, Guofei Gu

Texas A&M University
Amazon’s Alexa Is Totally Baffled by My Bilingual Family

Uber or Lyft with Google Home

Alexa skills top 80,000 after a big Alexa-powered holiday season

Hey Siri

amazon alexa

Google Assistant

Baidu 百度

Sony

Samsung

Bixby

mi
Do you really know/trust who (Voice Assistants or VAs) you are talking to?

Joey

Jennifer

Fortune
Voice Assistant

/ˈvaɪs əsɪstənt/ noun

• a computer program that can hold a conversation with somebody and complete particular tasks by responding to instructions or to information that it gathers from that person's digital device
• a computer program that can hold a conversation with somebody and complete particular tasks by responding to instructions or to information that it gathers from that person's digital device

Existing Work

• unrecognizable or inaudible malicious voice commands
  • Close Range
  • Physical
  • Speech Recognition
a computer program that can hold a conversation with somebody and complete particular tasks by responding **to instructions** or **to information that it gathers** from that person's digital device

Existing Work

- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition
• a computer program that can hold a conversation with somebody and **complete particular tasks** by responding **to instructions** or **to information that it gathers** from that person's digital device

Existing Work

• unrecognizable or inaudible malicious voice commands
  • Close Range
  • Physical
  • Speech Recognition

• speech misinterpretation
  • Remote
  • Application
  • Blackbox
a computer program that can hold a conversation with somebody and complete particular tasks by responding to instructions or to information that it gathers from that person's digital device

Existing Work

- recognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition

- speech misinterpretation
  - Remote
  - Application
  - Blackbox
- speech misinterpretation
  - Remote
  - Application
  - Blackbox
• speech misinterpretation
  • Remote
  • Application
  • Blackbox

Our Work
• unboxing the black box
• a systematically study of semantic misinterpretation (what’s after speech-to-text)
1. What’s in the Voice Assistant “blackbox”?

2. What’s the root cause of semantic misinterpretation?

3. How to explore the problem systematically?
Voice Assistant Application (vApp) Architecture

what’s inside the box

- Speech Recognition: from audio to text
- Semantic interpretation: from text to intents
- Audio Response: from text to audio
Our Work..

- Speech Recognition: from audio to text
- Semantic interpretation: from text to intents
- Audio Response: from text to audio
Semantic Misinterpretation

• Even when the speech recognition works fine, NLU yields incorrect intents.

Today in Google Home accidents I wanted to watch Nifty videos on YouTube and instead it played some really messed up “Patriotism for Kids” album on YouTube music. Oh no.

7:19 AM - 19 Jan 2019

1 Retweet 38 Likes

Jennifer
Semantic Misinterpretation

- Natural Language Understanding (NLU)
- Intent Classifier
Semantic Misinterpretation

Natural Language Understanding (NLU) → Intent Classifier → [Output]
Problematic Intent Classifier
root cause of semantic misinterpretation

Developer
Defined: “Alexa, install the true bank skill.”
Template:

“Alexa, tell me my balance.”

SYSTEM.InstallIntent
CUSTOM.BalanceQueryIntent
vApp Squatting

Victim vApp
The True Bank Skill

Malicious vApp:
The Truth Bank Skill

SYSTEM.InstallIntent

Root

install

skill

the true bank

truth

The True Bank Skill
The Truth Bank Skill
Fuzzing Semantic Misinterpretation
explore the problem systematically

• Fuzz potentially dangerous voice command variants
  • Problem: too many options! \(\Rightarrow\) large search space
Learn from Linguistic Knowledge

- Reduce the search space
- Find voice command utters that people likely to speak
- **Lapsus** is a concept beyond speech errors, it is anything that makes machine misunderstand your intents

- Regional language
- New Vocabulary
- Logical Fallacies
- ....
LipFuzzer

- Learn from existing linguistic knowledge
- Fuzz potentially dangerous voice command utters

- Pinpoint the problematic voice commands
- Help start to eliminate the threats
Evaluation for Lapsus and Lipfuzzer

• A misinterpreted voice commands can either due to ASR or the Intent Classifier
  • How much does the Intent Classifier (using developer defined templates) contribute?

• We collected 109 real Lapsus from Mturk (IRB approved)
• 77% of them are caused by improper intent classification

77% (84/109)

23% from ASR
Store-wide Scanning

- We use our trained Lapsus Model to find potentially vulnerable voice commands and vApps

- We let our chatbot speak with an Amazon Echo and a Google Home device
  - A verified vulnerable vApp is the one returns unintended result

<table>
<thead>
<tr>
<th>Store Name</th>
<th>Crawled vApp #</th>
<th>LipFuzzer-generated LAPSUS #</th>
<th>Potentially vulnerable vApp #</th>
<th>Verified vulnerable vApp % (Sampled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Alexa</td>
<td>32,892</td>
<td>497,059</td>
<td>26,790</td>
<td>71.5%</td>
</tr>
<tr>
<td>Google Assistant</td>
<td>2,328</td>
<td>11,390</td>
<td>2,295</td>
<td>29.5%</td>
</tr>
</tbody>
</table>
Attacking The True Bank Skill

• Use Amazon Mechanical Turk (MTurk) to let users “use” The True Bank Skill
  • We show the web user interface to MTurk workers
  • We collect their voice command audio records
  • Play these records to Alexa
• We check if users would eventually be using those generated “Malicious” vApp
• 4 Malicious ones are found to be “triggered”
Summary

• **Unboxes** the black box of VA platforms

• Pinpoint the **Intent Classifier** problem and **3-party vApp** development

• Find potentially dangerous voice commands in large scale (**LipFuzzer**)  

• An early signal that the lack of security considerations for vApp backend processing

• Limitations?
  • Need more data, more knowledge
  • Better Modelling and Machine Learning techniques
Take Away

• A similar trend: vApp and Mobile App

• More discovery even after the NLU component?

• How to deal with the human factors in Voice User Interface?

Automatic Speech Recognition (ASR) → Textual Data → Natural Language Understanding (NLU) → Intent Classifier → Intents

![Graph showing vApp and Mobile App growth over years]
Contact Us:
yangyong@tamu.edu

Project Release:
success.cse.tamu.edu/lab/lipfuzzer
Backup Slides
### Linguistic Knowledge Details

<table>
<thead>
<tr>
<th>Lapsus</th>
<th>Description</th>
<th>Examples</th>
<th>Example Logic Abstraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blends†</td>
<td>Two intended items fuse together when being considered.</td>
<td>Target: person/people</td>
<td>( \forall x, y, \text{phoneme}(\text{END}, &quot;S-N&quot;, x), \text{phoneme}(\text{END}, &quot;P-L&quot;, y) \rightarrow \text{phoneme}_\text{exch}(&quot;S-N&quot;, &quot;P-L&quot;, -) )</td>
</tr>
<tr>
<td>Morpheme Exchange</td>
<td>Morphemes changes places.</td>
<td>Target: He packed two trunks.</td>
<td>( \forall x, y, \text{suffix}(&quot;ed&quot;, x), \text{suffix}(&quot;s&quot;, y) \rightarrow \text{suffix}_\text{exch}(&quot;ed&quot;, &quot;s&quot;, -) )</td>
</tr>
<tr>
<td>Regional Vocabulary</td>
<td>Everyday words and expressions used in different dialect areas</td>
<td>Target: Entree</td>
<td>( \forall x, \text{word}(&quot;entree&quot;, x), \rightarrow \text{word}_\text{exch}(&quot;entree&quot;, &quot;hotdish&quot;, -) )</td>
</tr>
<tr>
<td>Category Approximation</td>
<td>Word substitution due to the lack of vocabulary knowledge.</td>
<td>Target: Show my yard camera.</td>
<td>( \forall x, \text{word}(&quot;show&quot;, x), \rightarrow \text{word}_\text{exch}(&quot;show&quot;, &quot;turn on&quot;, -) )</td>
</tr>
<tr>
<td>Portmanteaux‡</td>
<td>Combined words that are used.</td>
<td>Target: Eat the (late) breakfast</td>
<td>( \forall x, \text{word}(&quot;late breakfast&quot;, x), \rightarrow \text{word}_\text{exch}(&quot;late breakfast&quot;, &quot;brunch&quot;, -) )</td>
</tr>
</tbody>
</table>
BN Modeling
### Collected Lapsus Example

<table>
<thead>
<tr>
<th>Correct Form</th>
<th>LAPSUS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Installation Name</strong></td>
<td></td>
</tr>
<tr>
<td>“Thirty Two Money Tip with Nick True”</td>
<td>“Thirty Two Money Tip with Nick Truth”</td>
</tr>
<tr>
<td>”Elon - Tesla Car”</td>
<td>“Elon Tesla Car”</td>
</tr>
<tr>
<td><strong>Invocation Voice Command</strong></td>
<td></td>
</tr>
<tr>
<td>“Alexa, ask Elon to turn on the climate control”</td>
<td>“Alexa, ask Elon Musk to turn on the climate control”</td>
</tr>
<tr>
<td>“Alexa, ask massage manager begin session for number five”</td>
<td>“Alexa, ask massage messenger begin session for number five”</td>
</tr>
</tbody>
</table>
NLP Representation

Part-of-Speech:

Named Entity Recognition:

Basic Dependencies:

THE TRUE BANK SKILL
DH AHO . TR UW1 . BA E1 NG K . S K IM1 L .
# Paypal and Skyrim

<table>
<thead>
<tr>
<th>Intended Voice Command</th>
<th>LAPSUS</th>
<th>Effective LAPSUS?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Paypal&quot; (installation)</td>
<td>&quot;Pay-ple&quot;</td>
<td>✓</td>
</tr>
<tr>
<td>&quot;ask PayPal to check my balance&quot;</td>
<td>&quot;Pay-ples&quot;</td>
<td>✓</td>
</tr>
<tr>
<td>&quot;ask PayPal to check my balancing&quot;</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>&quot;ask PayPal to checks my balance&quot;</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>&quot;ask PayPal to checking my balance&quot;</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>&quot;Skyrim Very Special Edition&quot; (installation)</td>
<td>&quot;Skyrim Very Special Edit&quot;</td>
<td>✓</td>
</tr>
<tr>
<td>&quot;Skyrim Special Edition&quot;</td>
<td>&quot;Skyrim Special Edition&quot;</td>
<td>✓</td>
</tr>
<tr>
<td>&quot;Skyrim Very Specially Edition&quot;</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>&quot;Sky-ram Special Edition&quot;</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>&quot;Sky-im Special Edition&quot;</td>
<td>○</td>
<td></td>
</tr>
</tbody>
</table>

✓: Effective. ✗: Ineffective. ○: Maybe Effective