Fuzzing Semantic Misinterpretation for Voice Assistant Applications

Project Proposal

Presented by

Shivam Pandit Christin Wilson

Introduction

- **Security Concerns in Voice Assistants**
  - **Semantic Misinterpretations:** VA's are found to misclassify things based on speech understanding that can cause security concern
  - **Attacks on ASR & Intent Classifier:** Attacks on Automatic Speech Recognition and intent classifier are big concern as attackers leverage common spoken errors to breach vApp integrity for malicious intent

- **NLU's Intent Classifier**
  - Intent Classifier may misinterpret something other than user intent based on machine understanding

Introduction

- VUI based VA Architecture

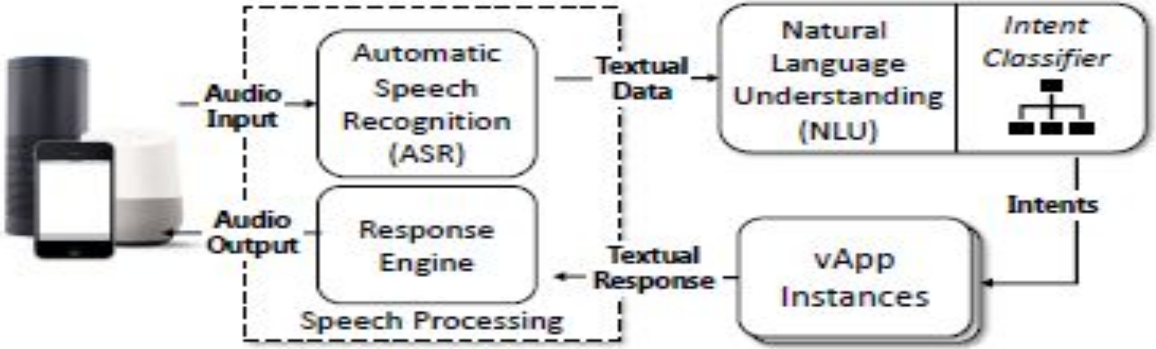
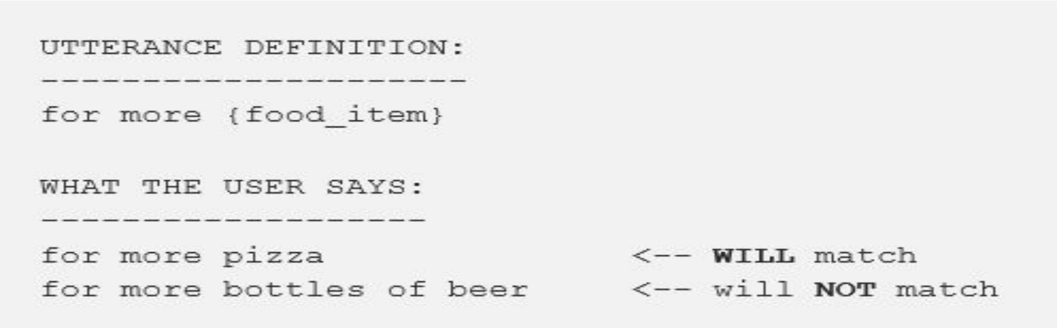


Fig. 1: VUI-based VA Architecture.

Introduction

- **Security Aspects**
  - Voice assistants allow us to directly control computational devices like phones, tablets that emphasizes need of security.
  - User voice commands can be misinterpreted by NLU intent classifier to give undesired results.
  - ASR and intent classifier are both proven to misinterpret the spoken command by users that can be leveraged by hackers to intrude privacy.
  - Developers can maliciously modify intent matching process in NLU.
  - Intent classifier plays more important role since it is last step of the interpretation process.

Introduction

- Examples
  - 

```
UTTERANCE DEFINITION:
---------------------
for more {food_item}

WHAT THE USER SAYS:
-------------------
for more pizza                   <-- WILL match
for more bottles of beer         <-- will NOT match
```

- The actual word used in the utterance matters as seen above.
- Machine learning algorithm considers the number of words in the utterance, the utterance word itself, as well as the number of words in each sample slot.

Attack Consequences

- Denial of Service
- Privacy Leakage
- Phishing
- Other consequences
    - Introduction of new functionalities like in-vApp purchasing can cause new consequences.

Related Work

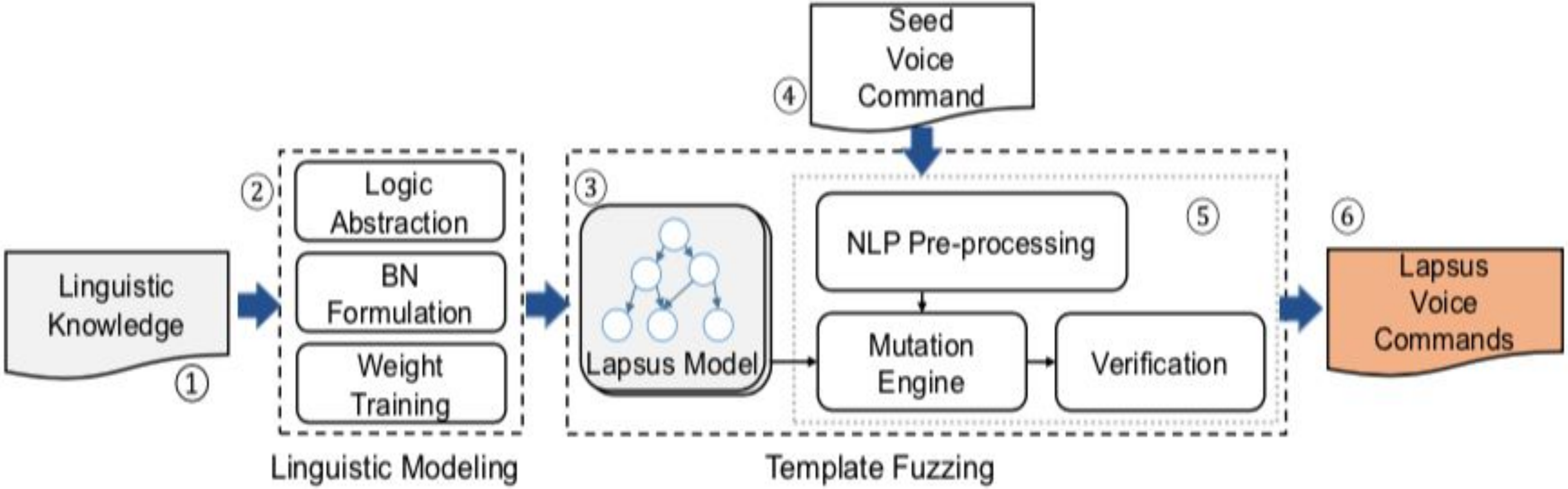
- Attacking ASR through Acoustic Channels
    - Launch Attacks that can be recognized by a computer speech recognition system but not easily understandable by humans.

- Attacking ASR with Misinterpretation.
    - vApp Squatting Attack
    - Uses a malicious skill with similarly pronounced name or paraphrased name to hijack the voice command meant for a different skill.

Implementation

1. **Creating BN Models**
   1.1. Linguistic Knowledge is collected and Bayesian Networks are formulated.
   1.2. Mispronunciation, Grammar and Vocabulary are considered.
2. **Collecting SEED inputs**
   2.1. Crawl skill commands from Alexa Skill Store and preprocess them for mutation.
3. **Perform Mutation**
4. **Evaluate**

Implementation



Timeline

Task	Date
Collect SEED inputs	2/25 - 3/4
Formulate BNs from collected Linguistic Knowledge	3/4 - 3/11
Train BNS with statistical weights and Preprocess SEED inputs	3/11 - 3/18
Perform Mutation	3/18 - 3/25
Prepare Midterm Project Prosentation	3/25 - 4/3
Evaluation	4/3 - 4/17
Prepare Final Project Presentation	4/17 - 4/24
Prepare Report	4/24 - 5/1

THANK YOU..