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

future tense

Amazon's Alexa Is Totally Baffled by My Bilingual Family

SMART HOME

Uber or Lyft with Google Home

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

Hey Siri

amazon alexa

Google Assistant

Patently Apple

Baidu 百度

SAMSUNG

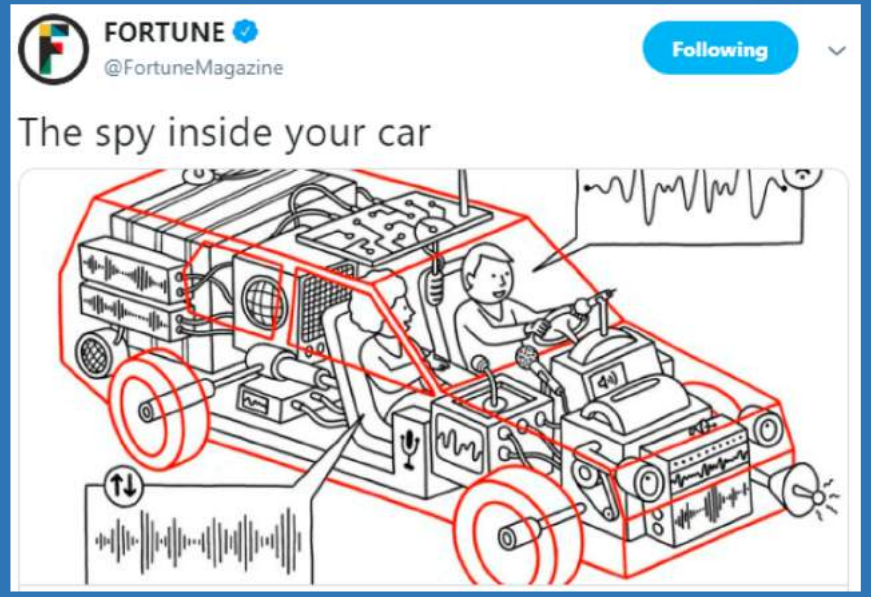
Bixby

SONY

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

Joey

Jennifer



Fortune

Voice Assistant

/ˈvɔɪs əsɪstənt/ OXFORD UNIVERSITY PRESS

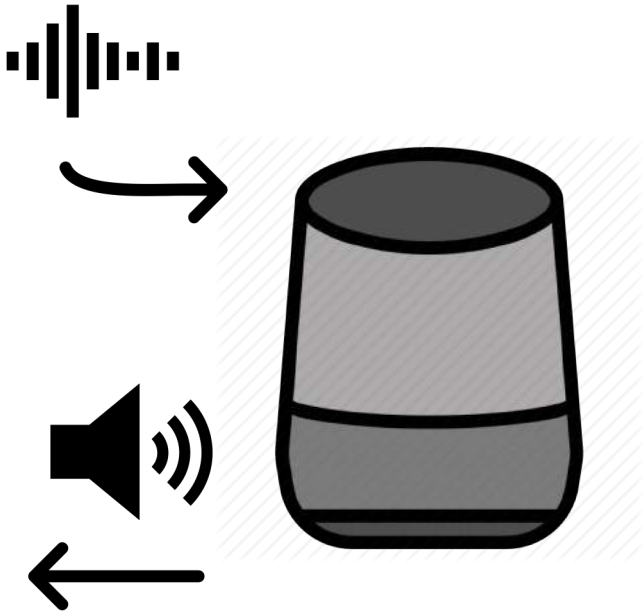
noun

- 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

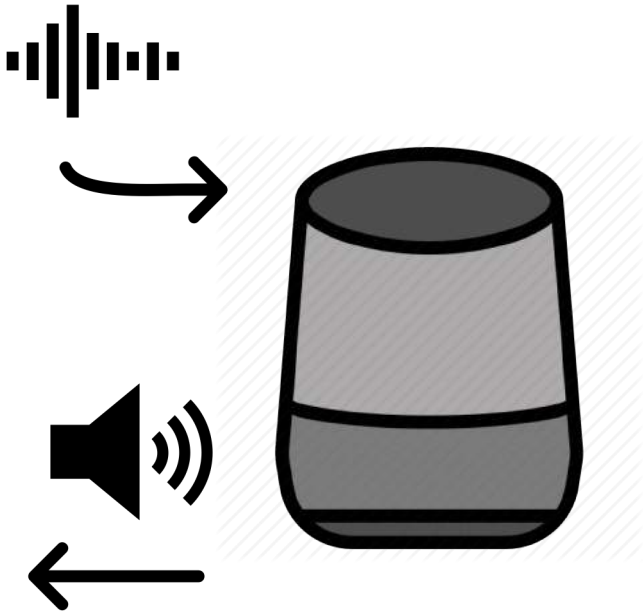
- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition



- 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

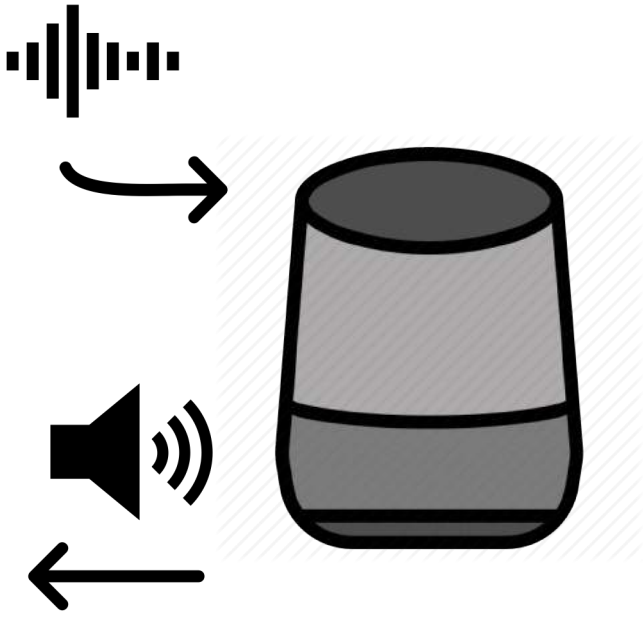
- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition



- 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

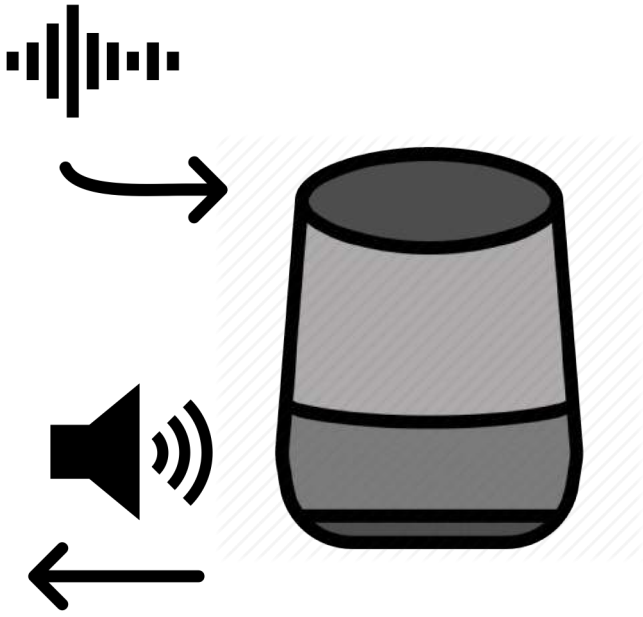
- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition
- speech misinterpretation
  - Remote
  - Application
  - Blackbox



- 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

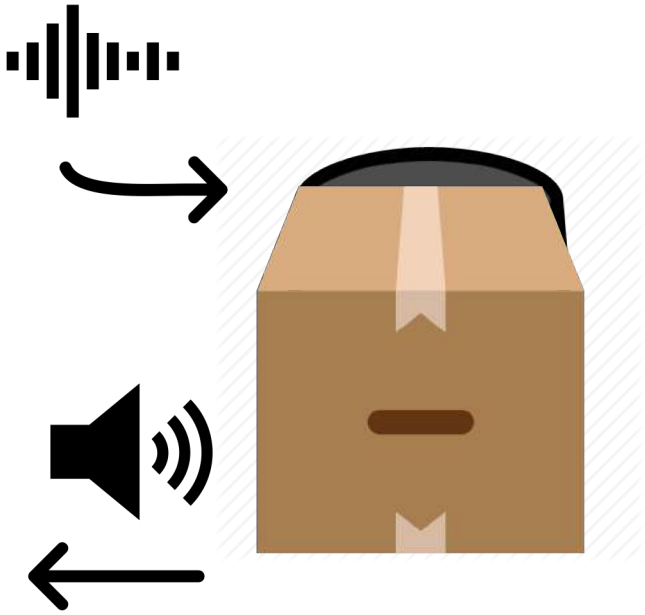
Existing Work

- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition
- speech misinterpretation
  - Remote
  - Application
  - Blackbox

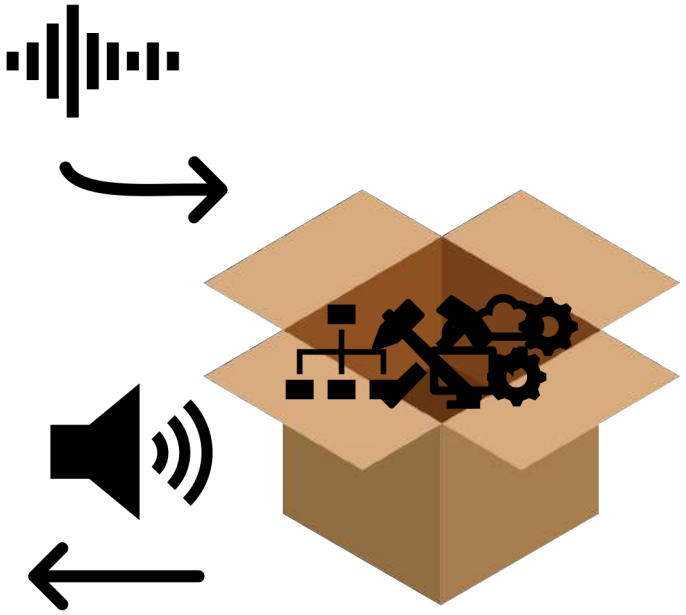


Our Work

- 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”?

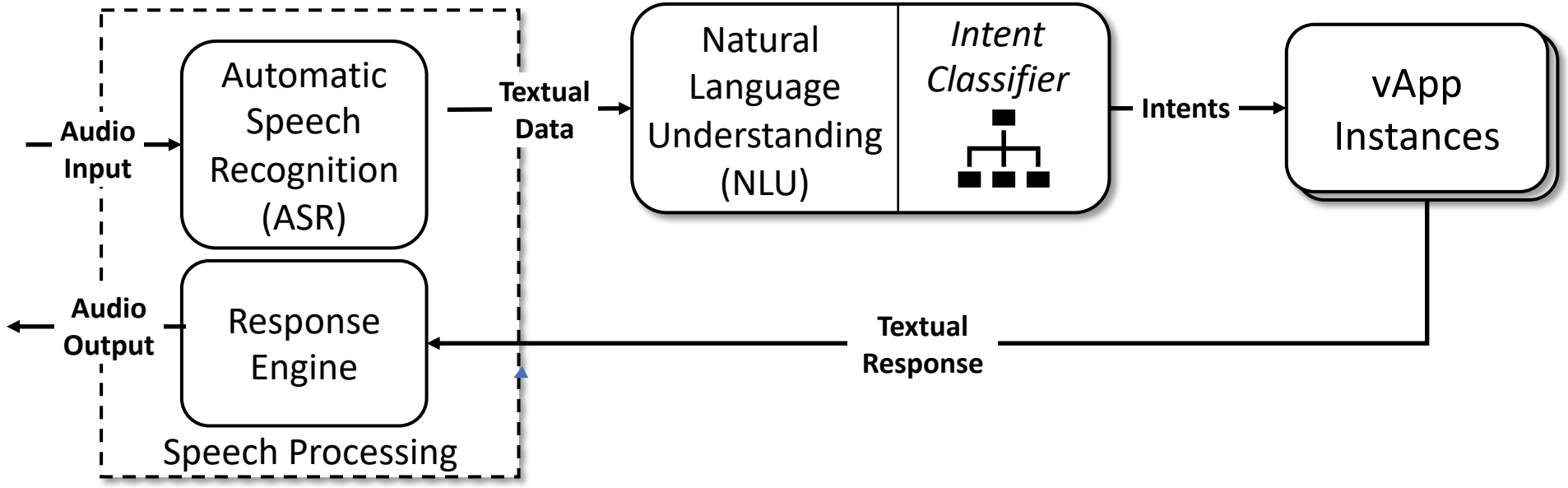
2. What’s the root cause of semantic misinterpretation?

3. How to explore the problem systematically?

Voice Assistant Application (vApp) Architecture

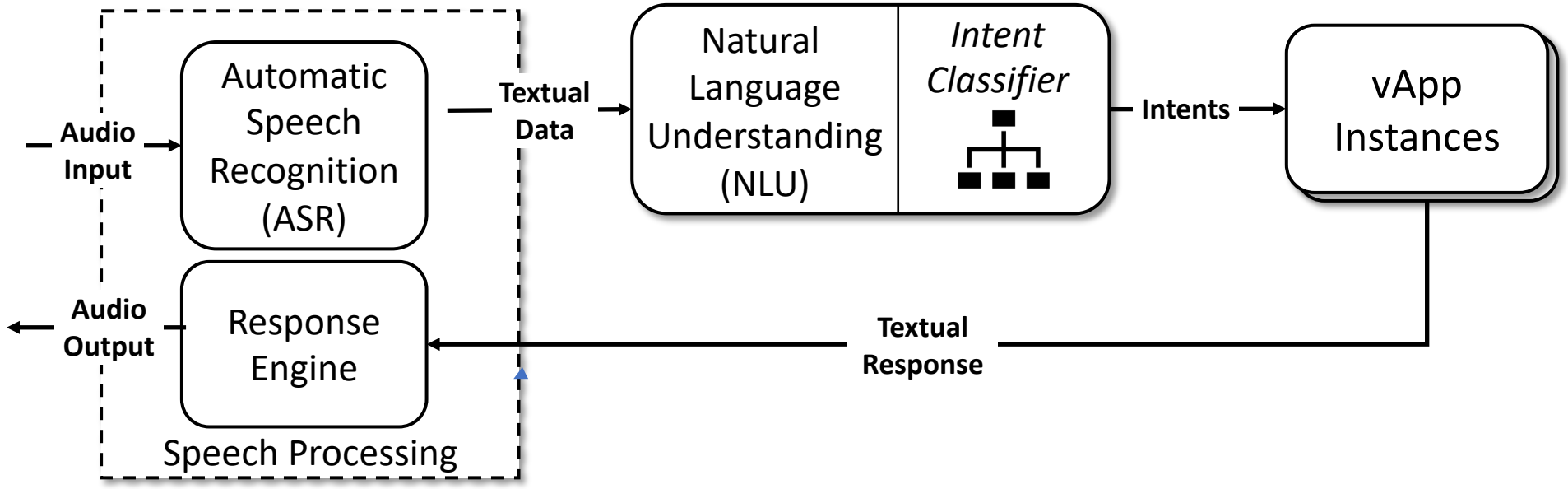
what's inside the box

- 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

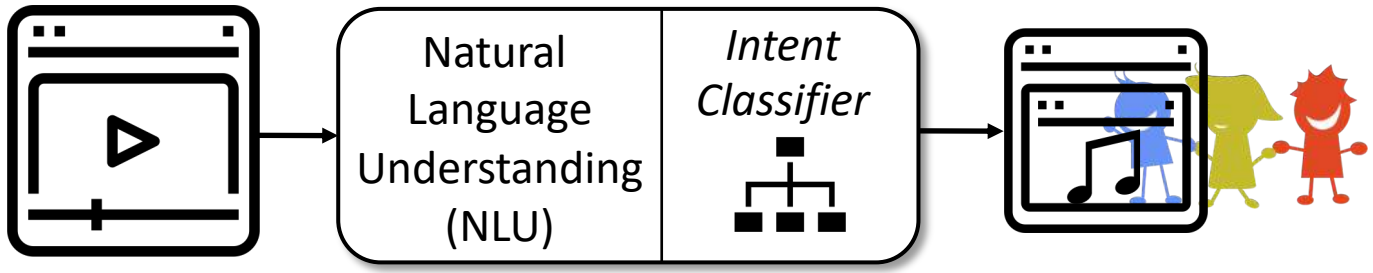


Semantic Misinterpretation

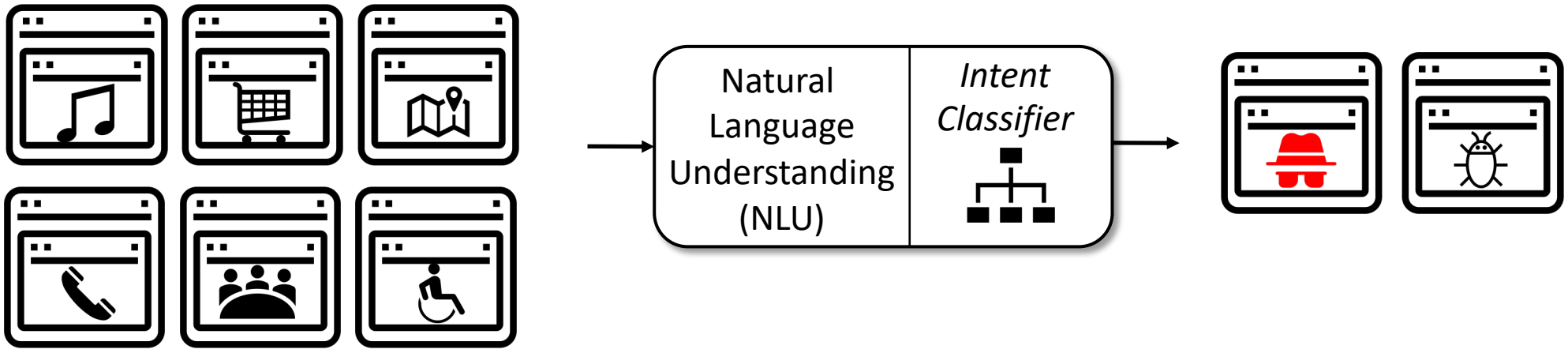
- Even when the speech recognition works fine, NLU yields incorrect intents.

Jennifer

Semantic Misinterpretation



Semantic Misinterpretation



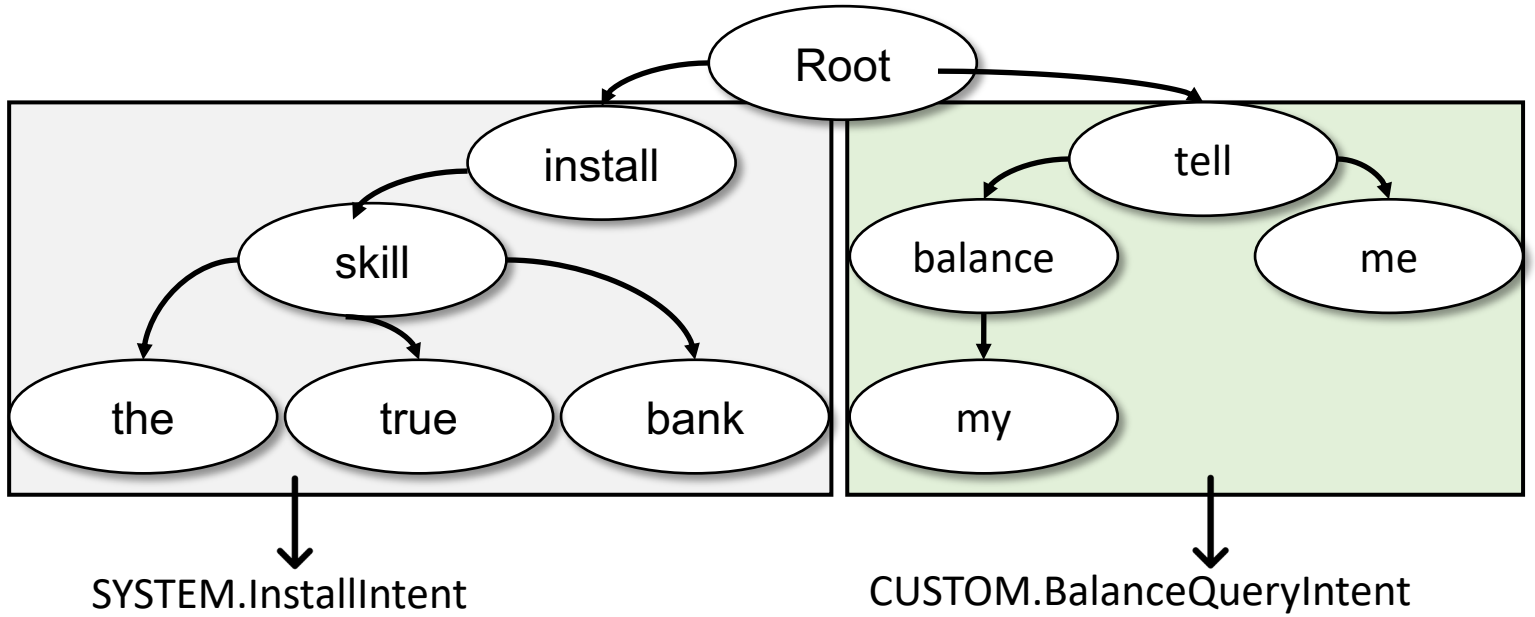
Problematic Intent Classifier

root cause of semantic misinterpretation

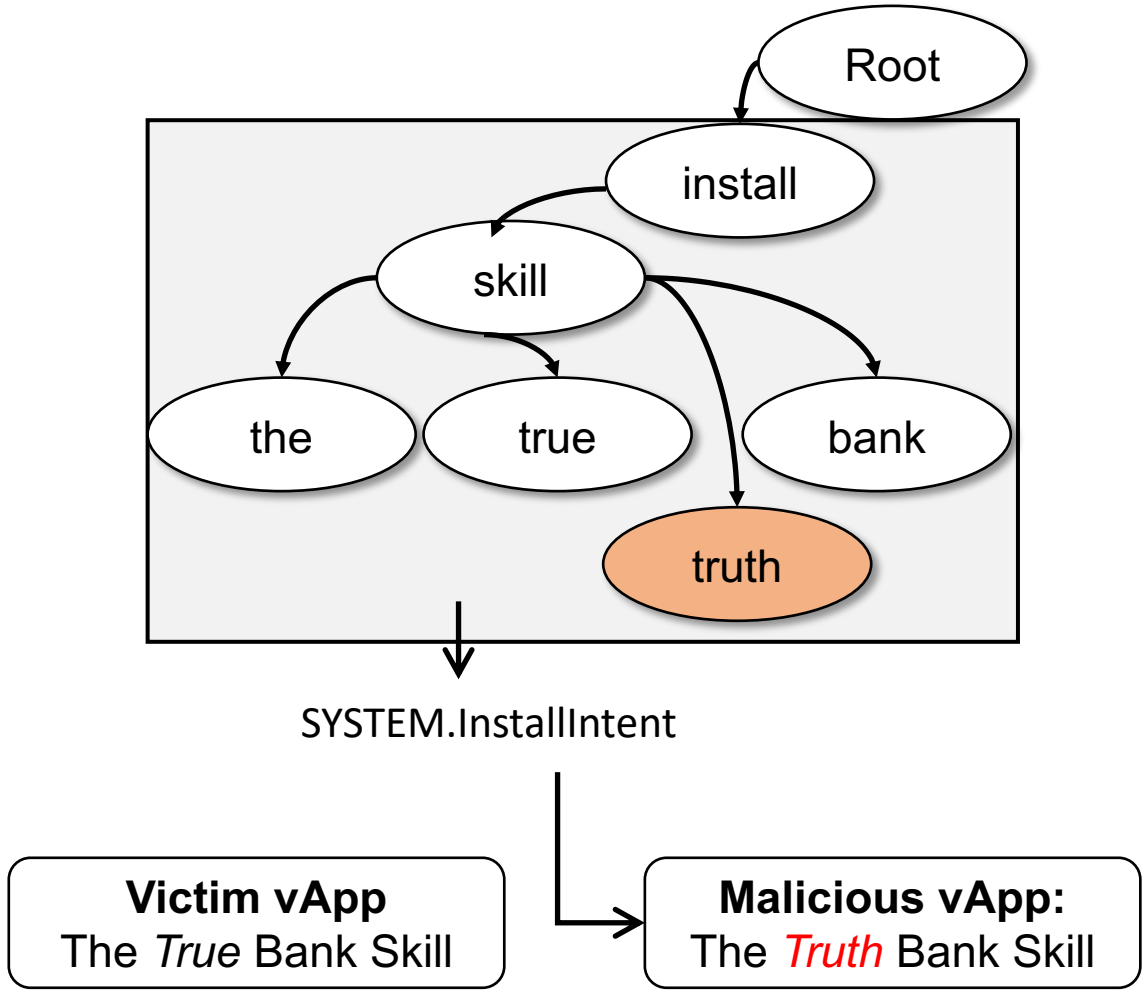
Developer Defined Template:

“Alexa, install the true bank skill.”

“Alexa, tell me my balance.”



vApp Squatting



Fuzzing Semantic Misinterpretation

explore the problem systematically

- Fuzz potentially dangerous voice command variants
  - Problem: too many options! ➔ large search space

LipFuzzer

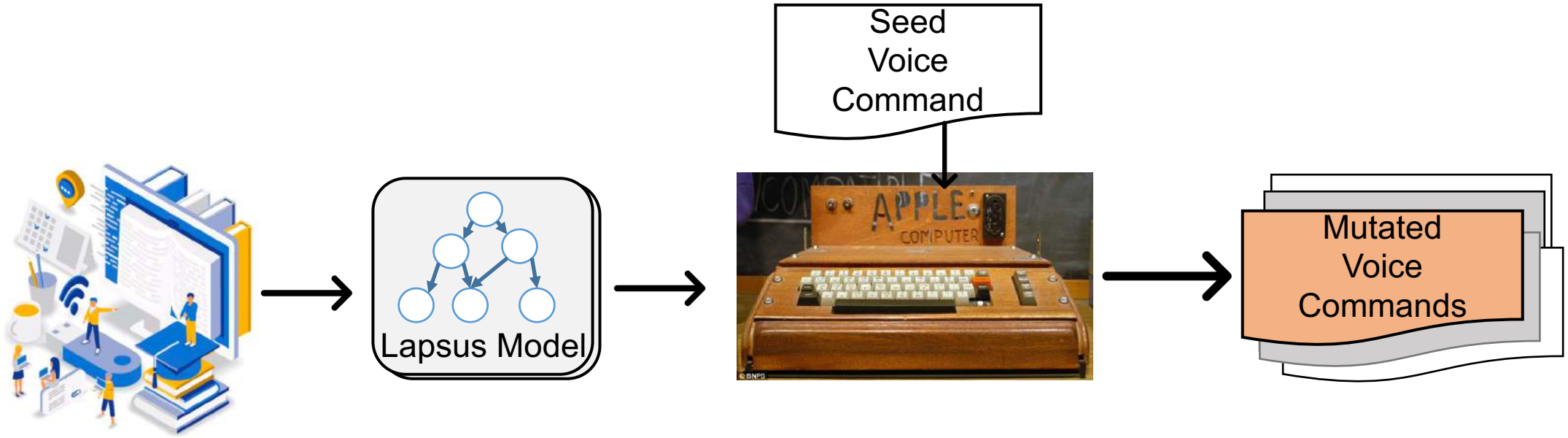
Learn from Linguistic Knowledge

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

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

LipFuzzer

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



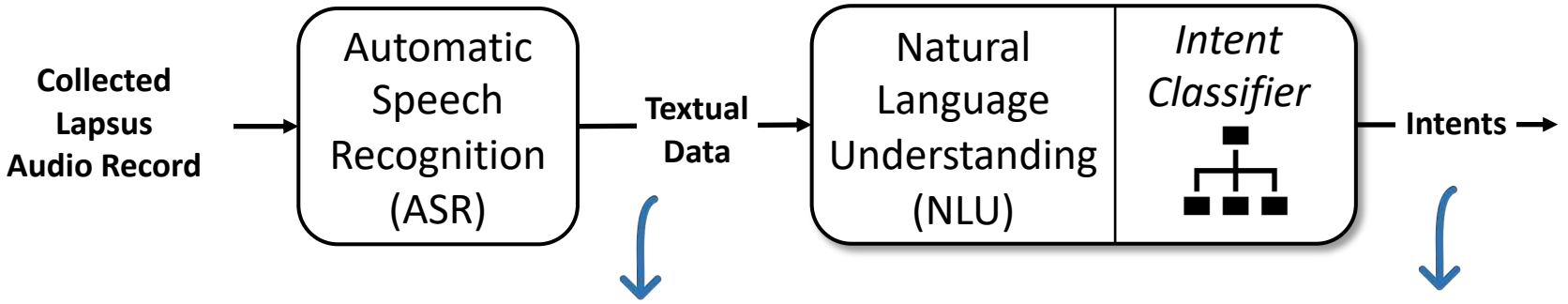
- Pinpoint the problematic voice commands
- Help start to eliminate the threats

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



Is text result transcribed correctly?

23% from ASR

Is the classification correct?

77% (84/109)

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

Store Name	Crawled vApp #	LipFuzzer-generated LAPSUS #	Potentially vulnerable vApp #	Verified vulnerable vApp % (Sampled)
Amazon Alexa	32,892	497,059	26,790	71.5%
Google Assistant	2,328	11,390	2,295	29.5%

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”

True Bank

View Skill ID

English (US)

Custom

2018-08-1

Summary

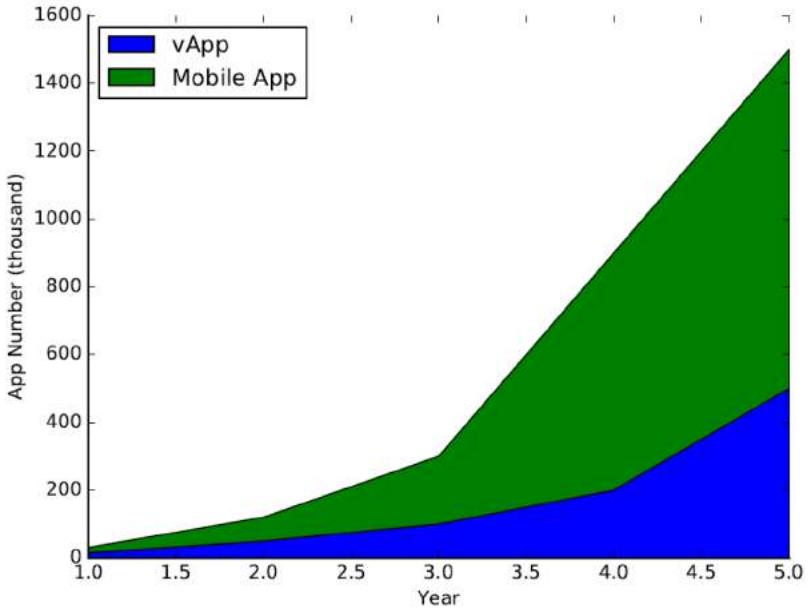
- **Unboxes** the black box of VA platforms
- Pinpoint the **Intent Classifier** problem **and 3-party vApp** development
- Find potentially dangerous voice commands in large scale (**LipFuzzer**)
- An early signal that the lack of security considerations for vApp backend processing

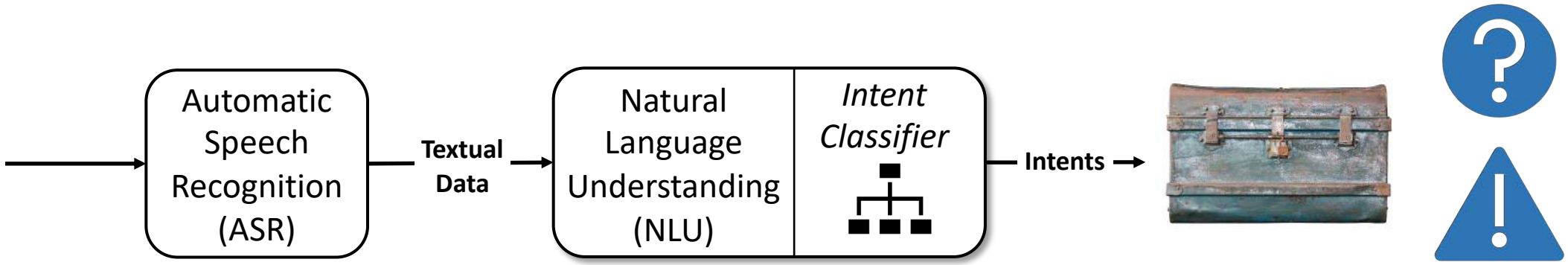
- Limitations?
  - Need more data, more knowledge
  - Better Modelling and Machine Learning techniques

Take Away

go beyond

- 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?





Contact Us:
yangyong@tamu.edu

Project Release:
success.cse.tamu.edu/lab/lipfuzzer

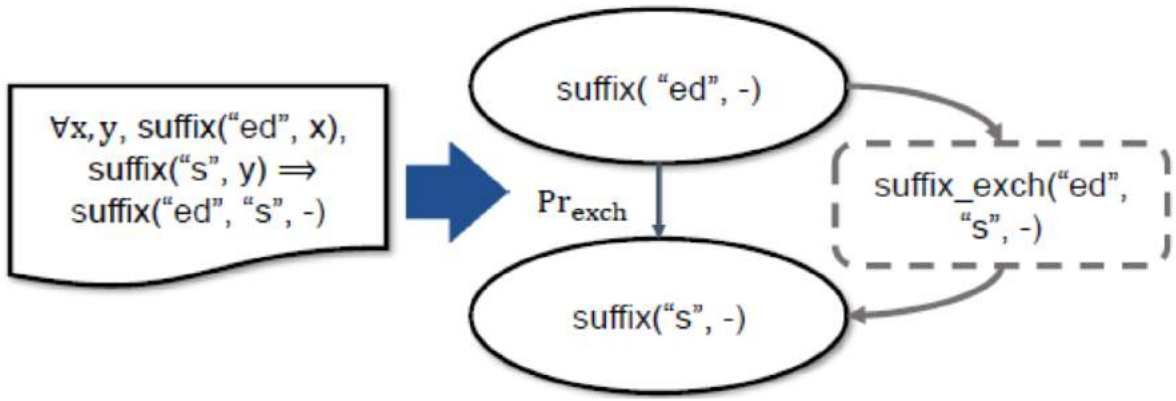
TEXAS A&M
UNIVERSITY

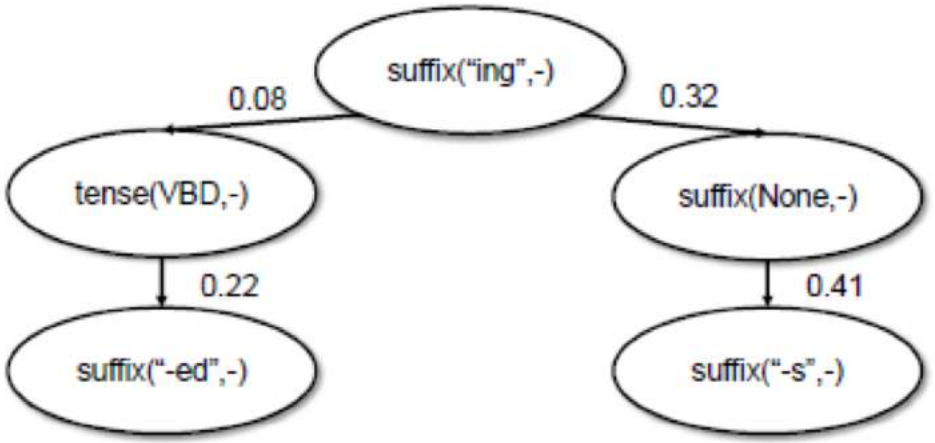
Backup Slides

Linguistic Knowledge Details

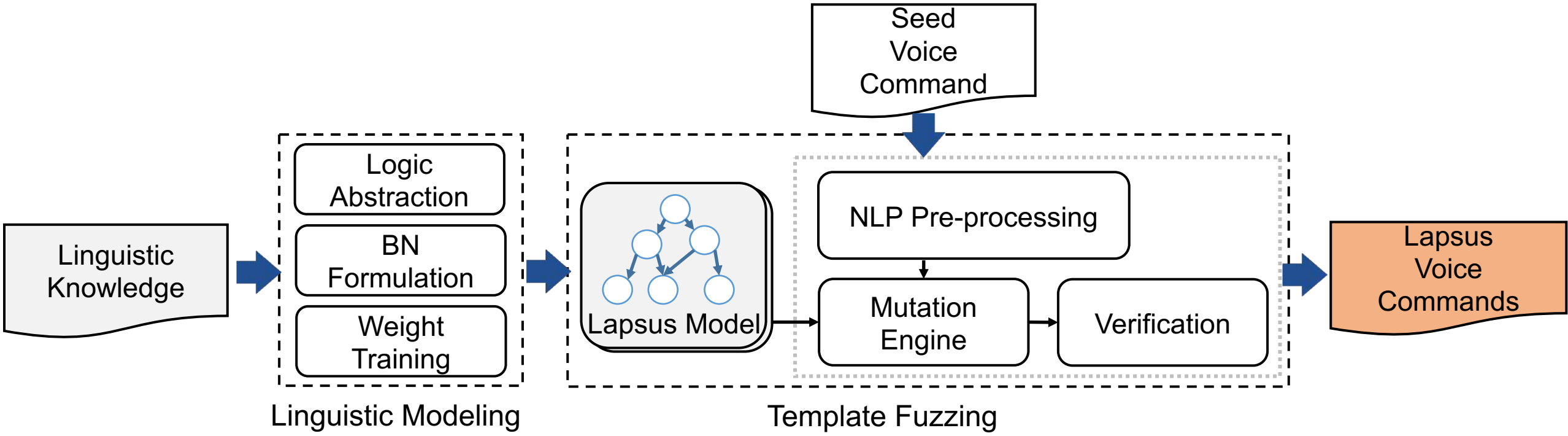
Lapsus	Description	Examples	Example Logic Abstraction
Blends†	Two intended items fuse together when being considered.	Target: person/people LAPSUS: perple	∀ x, y, phoneme(END, "S-N", x), phoneme(END, "P-L", y) → phoneme_exch("S-N", "P-L", -)
Morpheme* -Exchange	Morphemes changes places.	Target: He packed two trunks. LAPSUS: He packs two trunked.	∀ x, y, suffix("ed", x), suffix("s", y) → suffix_exch("ed", "s", -)
Regional ‡ Vocabulary	Everyday words and expressions used in different dialect areas	Target: Entree Lapsus: Hotdish (esp. Minnesota)	∀ x, word("entree", x), → word_exch("entree", "hotdish", -)
Category ‡ Approximation	Word substitution due to the lack of vocabulary knowledge.	Target: Show my yard camera. Lapsus: Turn on my yard camera.	∀ x, word("show", x), → word_exch("show", "turn on", -)
Portmanteaux‡	Combined words that are used.	Target: Eat the (late) brekfast Lapsus: Eat the brunch	∀ x, word("late breakfast", x), → word_exch("late breakfast", "brunch", -)

BN Modeling


∀x, y, suffix("ed", x), suffix("s", y) ⇒ suffix("ed", "s", -)
suffix("ed", -)
Pr_{exch}
suffix_exch("ed", "s", -)
suffix("s", -)


suffix("ing",-)
0.08
0.32
tense(VBD,-)
suffix(None,-)
0.22
0.41
suffix("-ed",-)
suffix("-s",-)

LipFuzzer


Seed
Voice
Command
Linguistic
Knowledge
Logic
Abstraction
BN
Formulation
Weight
Training
Linguistic Modeling
Lapsus Model
NLP Pre-processing
Mutation
Engine
Verification
Template Fuzzing
Lapsus
Voice
Commands

Collected Lapsus Example

	Correct Form	LAPSUS
Installation Name	"Airport Security Line Wait Times"	"Airport Security Wait for Line" "Airport Security Line **Waiting** Time" "Airport Line Wait Times"
	"Thirty Two Money Tip with Nick True"	"Thirty Two Money Tip with Nick **Truth**" "Thirty Two Money Tip with Nick **Drew**" "Thirty Two Money **Trip** with Nick **Truth**"
	"Elon - Tesla Car"	"Elon Tesla Car"
Invocation Voice Command	"Alexa, ask Elon to turn on the climate control"	"Alexa, ask Elon **Musk** to turn on the climate control"
	"Alexa, ask massage manager begin session for number five"	"Alexa, ask massage **messenger** begin session for number five"

NLP Representation

Part-of-Speech:

Named Entity Recognition:

Basic Dependencies:

- THE TRUE BANK SKILL
- DH AH0 . T R UW1 . B AE1 NG K . S K IH1 L .

Paypal and Skyrim

PayPal

by PayPal

Rated: Guidance Suggested

14

Free to Enable

"Alexa, ask PayPal to check my balance"

Skyrim Very Special Edition

by Bethesda Game Studios

Rated: Mature

242

Free to Enable

"Alexa, open Skyrim"

"Use shout"

Intended Voice Command	LAPSUS	Effective LAPSUS?
"Paypal" (installation)	"Pay-ple"	✓
	"Pay-ples"	✓
"ask PayPal to check my balance"	"ask PayPal to check my balances"	✗
	"ask PayPal to check my balancing"	✗
	"ask PayPal to check my balancing"	✗
	"ask PayPal to checks my balance"	✗
	"ask PayPal to checking my balance"	✗
"Skyrim Very Special Edition" (installation)	"Skyrim Very Special Edit"	✓
	"Skyrim Special Edition"	✓
	"Skyrim Very Specially Edition"	✗
	"Sky-ram Special Edition"	❍
	"Sky-im Special Edition"	❍

✓: Effective, ✗: Ineffective, ❍: Maybe Effective