AI's Multilingual Journey in Search and Generation of Cross-Language Sound-squatting

Rodolfo Vieira Valentim
Supervisor: Prof. Marco Mellia

Research context and motivation

- One common attack is domain squatting, which occurs when attackers register perceptively confusing domain names aiming at tricking visitors into them.
- Sound-squatting has gaining traction with the advent of smart speakers and voice-assistants.
- The state-of-art in detection uses statically built lists of homophones.
- We hypothesize Artificial Intelligence can produce more comprehensive sound-squatting candidates and be used as an automatic method for sound-squatting generation and detection.

Addressed research questions/problems

- Harnessing AI for Proactive Generation: How can AI-driven techniques be leveraged in sound-squatting?
- Evaluating Quality and Scalability: How can we explore the quality and scalability of generating multiple sound-squatting candidates from a single target?
- Crossing Language Boundaries: How to generated homophones for different languages?

Novel contributions

- Introducing a Methodology: Methodology for generating sound-squatting candidates across multiple languages.
- Assessing Domain Abuse: Analyzing the existence of sound-squatting abuse in domain registration.
- Probing PyPI Abuse: Investigating sound-squatting abuse in the Python Package Index (PyPI).
- Benchmarking Against the Best: Comparing sound-squatting with state-of-the-art squatting techniques.

Adopted methodologies

- Improve Phonemic Representation: Utilize acoustic features of phoneme tokens to handle untrained languages.
- Improve Generation: Top-p beam search to produce multiple candidates.
- Enable Multilanguage: Target language as initial state for multi and cross-language generation.
- Consistent Validation: Validate using known homophonic coverage.

Results

- Sound-Squatting in Domain Names:
  - Data: Utilize a dataset of more than 900 million certificates collected on CT logs.
  - Findings: Approximately 1.1 million possibly registered domains for 10,000 targets, impacting around 95.8% of them.

- Sound-Squatting in Python Packages:
  - Data: Analyze candidates generated over 900 days.
  - Findings: 1,579 registered candidates potentially squatting 951 existing packages out of 5,000.

Future work

- Enhanced Sound-Squatting Detection: Develop more robust detection mechanisms for sound-squatting, considering evolving techniques and tactics used by cyber attackers.
- Behavioral Analysis: Investigate the behavioral patterns of sound-squatting in various contexts (e.g., domain names, social media usernames) to refine detection and mitigation strategies.
- User Education: Explore ways to educate users and organizations about the risks associated with sound-squatting and other emerging cybersecurity threats.

Publications

- Published works: 1 journal and 3 conference workshop
- Submitted works: 1 journal