Description for Artifact of "Measuring the Effectiveness of Privacy Policies for Voice Assistant Applications"

September 10, 2020

Abstract

This document contains a description of the artifact components and detailed instructions on how to download and run the artifact's software and data. All our data are available at https://github.com/voice-assistant-research/voice-assistant

1 Introduction

Our artifact consists of three parts. 1) Our dataset ("dataset" folder), which includes all the skills and actions we collected and the privacy policies of these skills and actions. This part corresponds to Section 4.1 "High-level Issues" in our paper. 2) Our code, which analyzes the privacy policy and description ("code" folder). Three python scripts are used to analyze the data practice in each privacy policy, detect inconsistency between the privacy policy and description, and find skills lacking a privacy policy. 3) The result of our paper, which is in the folder "results". The second and third parts are about Section 4.2 in our paper. Our code was developed and should be run on a Linux system.

2 Dataset

There are 8 files in the "dataset" folder. This part corresponds to Section 4.1 "High-level Issues" in our paper.

"1_all_skills_64720.zip" includes 64720 skills we collected with the skill id, skill name, category, developer, rating number, rating star, invocation phrase, description and privacy policy.

"2_all_actions_2201.csv" includes 2201 actions we collected with action category, link, name, developer, rating star and privacy policy.

"3_skills_with_policy_17952_with_duplicate.xlsx" includes 17952 skills with their privacy policy links. We also list the skills with duplicate links on top and with red color (by using "highlight duplicate values function" in office). 10124 skills use a duplicate link.

"4_actions_with_policy_1967_with_duplicate.xlsx" includes 1967 actions with privacy policy links. 239 actions use a duplicate link.

"5_skills_privacy_policy_content" compresses 14764 skills' privacy policies in txt format and "6_actions_privacy_policy_content" includes 1909 actions' privacy policies. They would be used if you want to analyze all the skills and actions.

"7_official_skills.xlsx" and "8_official_actions.xlsx" includes 98 skills developed by Amazon and 110 actions developed by Google as well as their privacy policy information.

3 Code

The second part is our code to analyze the privacy policy and description ("code" folder).

For running the code, one need to install a python library named "spacy" first (with "pip install spacy" under python3). Then you need to uncompress the file "en_core_web_sm-2.2.5.zip", which is the data for the library. (You can uncompress all the compressed files in "dataset" and "code" folders. All these files would be used.) You should run the script under the "code" folder.

Since there are a large number of skills and actions, analyzing all of them would cost a long time (We run all the code on a computer with 6 cores i5-9500 CPU and 16 GB memory for tens of hours), we provide you 100 skills for testing. You can directly uncompress the "example_skills_privacy_policy.zip" and the 100 privacy policies would be tested. If you want to test all the skills and actions, you can copy the "5_skills_privacy_policy_content" and "6_actions_privacy_policy_content" from "dataset" folder to the "code" folder and uncompress them. All of them would include a folder named "privacy_policy" and you do not need to edit the code in the python file.

3.1 Data Practice

"1_get_data_practice.py" can be used to get the data practice for each privacy policy. Use the command "python3 1_get_data_practice.py" to run the code and it would generate a file named "1_result_data_practice.txt", which contains the skill id (start with "B") and the number of data practice in the privacy policy of the skill. You can test the "example_skills_privacy_policy.zip" to test 100 skills, "5_skills_privacy_policy_content" to test all skills or "6_actions_privacy_policy_content" to test all actions.

3.2 Inconsistency Between Privacy Policy and Description

"2_get_inconsistent_skill.py" can detect the inconsistency between skill description and privacy policy. Uncompress the "1_all_skills_64720.csv.zip" in "dataset" and use the command "python3 2_get_inconsistent_skill.py" to run the code and it would generate a file named "2_result_inconsistent_skill.txt". In the file, if a skill doesn't have an inconsistency, it shows skill id and "0"; if it has, it shows the skill id and the phrase in the description which is inconsistent with privacy policy (such as "we need your email", but "email" is not mentioned in the privacy policy). We didn't find any inconsistency for Google actions, so you can test 100 example skills or all the skills.

3.3 Missing Required Privacy Policy

"3_lack_privacy_policy.py" can detect whether a skill lacks a privacy policy. For testing the code, uncompress the "1_all_skills_64720.csv.zip" in "dataset" and run the "python3 3_lack_privacy_policy.py". It would generate a file named "3_result_lack _privacy_policy.txt". If a skill lacks the privacy policy, it shows skill id and the data practice in description (such as "we need your email"); otherwise, it shows skill id and "0". Since this code would analyze all the skills without privacy policy (46768 skills), it costs a long time and you can stop the program at any time to check the result of the analyzed skills. We didn't find any google action lacking privacy policy, so you could only test all the skills. (Move "privacy_policy" from "dataset" to "code". We check all skills without a privacy policy. If you use 100 skills, the script would treat all other skills as no privacy policy and generate lots of wrong results.)

4 Results

The third part is the result of our paper, which is in the folder "results".

"1_skill_data_practice.txt" includes the number of data practice in privacy policy for all skills.

"2_action_data_practice.txt" includes the number of data practice in privacy policy for all actions.

"3_inconsistent_skills.xlsx" shows the 50 skills we detected whose privacy policy is inconsistent with the description. (Section 4.2.3 in the paper)

"4_lack_privacy_policy_skills.txt" shows the 20 skills which lack a privacy policy. (Section 4.2.4 in the paper)

"5_cross_platform_apps_with_different_link.txt" provides the voice-apps present on two platforms but have different privacy policies. 13 skill & action pairs have different privacy policies and 10 skill & action pairs have no privacy policy for Amazon skills.