Avtomatizacija večpredstavnostnih ukazov s pomočo storitve glasovnega nadzora preko Google Assistant in Wit.ai

Hrvoje Zeman, Aleš Zamuda
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46, 2000 Maribor
hrvoje.zeman@student.um.si, ales.zamuda@um.si

Automating Multimedia Commands over a Voice Control Service through Google Now and Wit.ai

This contribution for the Student section describes the process of creating a virtual assistant inspired by Google's Google Now Assistant or Amazon's Alexa, and describes the technologies used a service to process natural language, which is the main mode of communication. The program uses Google's speech-to-text technology to receive a voice command and uses Facebook's Wit.ai service to determine what the user wants. We used the Tkinter library to display the results and create a simple graphical interface.

The result is the creation of a virtual assistant instructed by the user through speech and this is able to perform automatization of simple daily tasks like hands-free multimedia commands.

Kratek pregled prispevka

Ta prispevek za študentsko sekcijo opisuje postopek ustvarjanja navideznega pomočnika, navdihnenjega s strani Googlovega Asistenta Google Now ter Amazonove Alexe in opisuje tehnologije, ki se uporabljajo kot storitev za obdelavo naravnega jezika, ki je glavni način komunikacije. Program uporablja Googlovo tehnologijo pretvorbe zvoka v besedilo za sprejem glasovnega ukaza in s pomočjo Facebookove storitve Wit.ai določa, kaj želi uporabnik. Za prikaz rezultatov in ustvarjanje preprostega grafičnega vmesnika smo uporabili knjižnico Tkinter.

Končni rezultat je ustvarjanje navideznega asistenta, upravljanega s strani uporabnika preko govora, kar nudi avtomatizacijo opravljanja preprostih dnevnih opravil, kot so prostoročni multimedijski ukazi.

1 Uvod

Umetna inteligenca (UI) je področje računalniških znanosti, ki se ukvarja z razvojem sistemov, ki naj bi bili sposobni opravljati naloge za katere je bila potrebna človekova inteligenca, kot npr. razpoznavanje govora, vizualna percepcija, avtonomna vožnja, odločanje itn. [1]

Čeprav že dolgo obstajajo tehnologije, ki zmanjšujejo težo in povečajo intuitivnostodnosa med človekom in računalnikom, so bile tehnologije doslej kombinirane le v nekaj aplikacijah [2]. Tako vidimo primer uporabe tehnologij v virtualnih inteligentnih pomočnikih, kot je Google Now [3], manj znani primeri uporabe tehnologije pa so glasovno iskanje v brskalniku Google Chrome. Sistemi pretvorbe govora v besedilo so bili razvijani že za časa programa Audrey [4], ki je lahko govor pretvoril v besedilo, vendar je bil omejen na razumevanje števil. Leta 1960 se je pojavil računalnik ELIZA, ki je z uporabo metodologije prepoznavanja in nadomeščanja vzorcev lahko vzdrževal pravi pogovor, kar je ustvarilo iluzijo, da program razume človeka.

McLean je napovedal, da bodo inteligentni virtualni pomočniki v prihodnosti za vsakodnevne dejavnosti nadomestili tehnologijo, kot so namizni in prenosni računalniki. [5]

V svojem prispevku, ki obravnava vpliv virtualnih asistentov na vsakodnevno poučevanje v učilnicah na novozelandskih šolah, Butler navaja, da tehnologija hitro spreminja način medsebojne komunikacije učencev in da digitalna doba pomembno vpliva na to spremembo. [6]

V medicini je bila umetna inteligenca uporabljena za zgodnjo detekcijo epilepsije na področju nevrologije in za oceno kardiovaskularnega tveganja na področju kardiologije [7] ali npr. za določanje korelacije med diagnozami [8].

V naslednji sekciji so podana sorodna dela. V tretji sekciji sledi opis razvite metode in v četrti njena demonstracija. V zadnji sekciji so podani sklepi in predlogi za nadaljnje delo.

2 Sorodna dela

Deuerlein in Langer sta se svojem delu ukvarjala s prevajanjem zvoka v besedilo. Sistem opisuje postopek razvoja programskega vmesnika, ki uporablja oblak za prepoznavanje uporabniškega ukaza in nato izvajanje določene naloge. Avtorji opišejo tudi dve kategoriji, v katere je mogoče razvrstiti avtomatizirano komunikacijo človek-robot: vprašanja in naloge. [9]

Rahman, se je odločil, da je bolje oblikovati programe za pogovor z ljudmi z vzorcem, napisanim v AIML, čeprav je to le eden od pristopov, ki ga predlaga. [10].

Poleg tega P. Singh v svojem prispevku opisuje, zakaj je v računalnikih izjemno težko proizvesti zdravo pamet, in opisuje načine, kako bi to lahko dosegli. [11]

3 Metoda

Po uvozu potrebnih knjižnic, ki se ukvarjajo z različnimi funkcionalnostmi, kot so komunikacija z operacijskim sistemom, komunikacija z internetnim brskalnikom, se zažene program, ki služi uporabniku.

Nato program preveri, ali v računalniku obstaja uporabniški ID, da lahko preveri, ali ima Google Račun podporo za uporabo storitve pretvorbe govora v besedilo. Če datoteka ne obstaja, se program konča.

Po zagonu programa lahko uporabnik začne govoriti in daje ukaze programu, kar pomeni, da je glavni način komunikacije med programom in uporabnikom glas.

Med uporabnikovim govorom program zbere dele zvoka v lokalni vsebnik, in ko zbere dovolj podatkov, jih program pošlje Googlovi storitvi v oblaku, ki pretvori zvok v besedilo, in to besedilo se vrne kot rezultat klica zunanje storitve z uporabo API-ja REST.

Ko Googlov API vrne besedilo, ki ga je izgovoril uporabnik, se besedilo pošlje drugi zunanji storitvi, Wit.ai Speech API, ki se ukvarja z ugotavljanjem namena besedila.

Potem ko je drugi API vrnil rezultate, imamo zdaj podatkovne ukaze, ki nam povedo, kaj je uporabnik želel narediti. Naslednji korak je obdelava standardne izhodne oblike API-ja REST, ki je v našem primeru zapis JSON.

V programskem jeziku Python za branje niza JSON ni potrebna dodatna knjižnica, kar ima jezik sam vgrajeno zmožnost branja nizov JSON.

Po branju vrednosti niza JSON, ki izraža, kaj je uporabnik želel narediti, lahko zdaj te podatke primerjamo z vnaprej določenimi rezultati, zapisanimi v programski kodi. Tako lahko preverimo, kateri nabor podatkov se ujema s katerim ukazom v programski kodi, in na podlagi tega preverjanja program sproži določena dejanja, ki so skladna z zahtevami uporabnika.

Nato se uporabniku prikaže določeno besedilo s pomočjo knjižnice Tkinter v skladu z izvršenim ukazom. Če je bil ukaz iskanje informacij o osebi, dogodku ali lokaciji, zaradi česar je besedilo preveliko za tiskanje s knjižnico Tkinter, se besedilo shrani v zunanjo datoteko, ki se nato odpre in prikaže uporabniku.

Če na primer uporabnik reče »Show me images of New York«, kot vidimo na sliki 1, bo program obdelal ukaz in v Googlu Images odprl rezultat iskanja New Yorka.

Na sliki 2 vidimo del programske kode, napisane v obliki psevdokode.

```
Find images of New York.
```

Slika 1: Postopek iskanja slik o New Yorku

Celotna koda je na voljo na naslednjem naslovu: https://github.com/Zemaaan/Red-Queen/tree/main

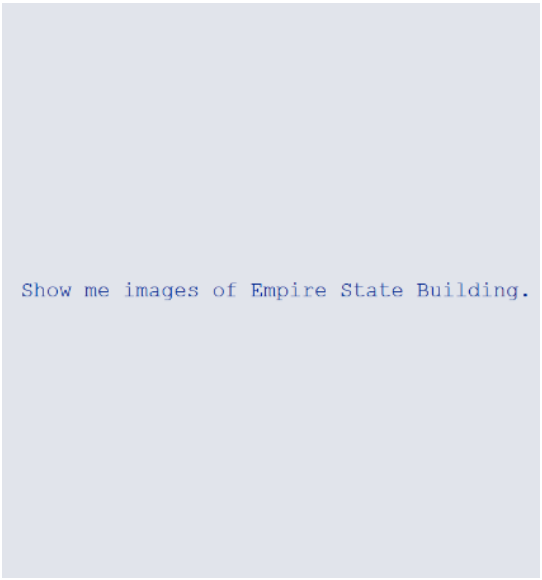
Vrstici 334 in 335 obravnavata definiranje predmetov, ki se ukvarjajo z vizualizacijo grafičnega vmesnika, in pretvorbo glasu v besedilo. Vrstici 336 in 337 obravnavata tekoče niti, ki bodo nadzirale grafični vmesnik in pretvarjanje glasu v besedilo. Vrstica 573 se ukvarja z obdelavo povratnih informacij iz API-ja Wit.ai Speech za obdelavo naravnega jezika in iz nje prebere uporabnikovo željo. Vrstice 575 - 582 se ukvarjajo s preverjanjem, ali namerava uporabnik iskati informacije, in če je tako, prebere zahtevani izraz iz podatkov JSON, poišče informacije z uporabo modula Wikipedia in te podatke zapiše v datoteko in odpre datoteko za uporabnik. Vrstica 583 se ukvarja s preverjanjem, ali namerava uporabnik zakleniti računalnik. Vrstica 584 obvesti uporabnika o dejanju. Naslednja vrstica doda nit na seznam aktivnih niti. Vrstica 587 zaklene računalnik. Vrstica 588 preveri, ali namerava uporabnik odštevati. Vrstica 589 zahteva število, od katerega se odšteva. Vrstica 591 kode pa odšteja določenega števila do nič. Vrstica 595 obvesti uporabnika, da bo program odprl kamero. Vrstica 597 doda nit na seznam aktivnih niti in vrstica 598 odpre kamero. Vrstice 618 - 621 se ukvarjajo z nadzorom temnega načina. Vrstica 622 se ukvarja s preverjanjem, ali namerava uporabnik iskati slike na določeno temo. Vrstica 624 išče izraz za iskanje slik v podatkih JSON. Vrstica 627 odpre Google Slike za določen termin.

4 Rezultati

V tem delu bomo razložili rezultate realiziranega in pokazali, kako uporabiti program, ki smo ga realizirali. Program nadzoruje predvsem glas, kar močno poenostavi nekatere vsakdanje postopke, na primer odpiranje datoteke, zaklepanje zaslona ali brskanje po internetu in odpiranje kamere.

```
334.  GUIThread = MainWIndow()
335.  TTSThread = SpeechToText(GUIThread)
336.  start thread GUIThread
337.  start thread TTSThread
573.  UserIntent = API output intent
575.  if UserIntent je FindInformation:
          577.  Find search term in JSON response
          578.  Use Wikipedia module to find information
      579.- 581.  Write the information to a file
          582.  Open file for a user
583.  if UserIntent == LockComputer:
          584.  Notify user
          586.  Add thread to a list of active threads
          587.  Lock the computer
588.  if UserIntent == CountDown:
          589.  Search countdown number in JSON response
          591.  Countdown
594.  if UserIntent == OpenCamera:
          595.  Notify user
          597.  Add thread to a list of active threads
          598.  Open a camera
618.  if UserIntent == TurnOnDarkMode:
          619.  Turn dark mode on
620.  if UserIntent == TurnOffDarkMode:
          621.  Turn dark mode off
622.  if UserIntent je SearchImages:
          624.  Find search term in JSON response
          627.  Find images for search term on the internet
```

Slika 2: psevdokod dela aplikacije

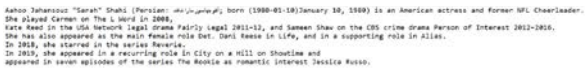


Slika 3: Postopek odpiranja Google Images iskanja za Empire State building.

Če uporabnik izda ukaz za iskanje informacij o sobi osebi ali lokaciji, bo program izbral zahtevani izraz in modul Wikipedia uporabil za iskanje zahtevanih podatkov v Wikipediji in te podatke prikazal uporabniku.

Slika 4: Postopek iskanja informacij o igralki Sarah Shahi.

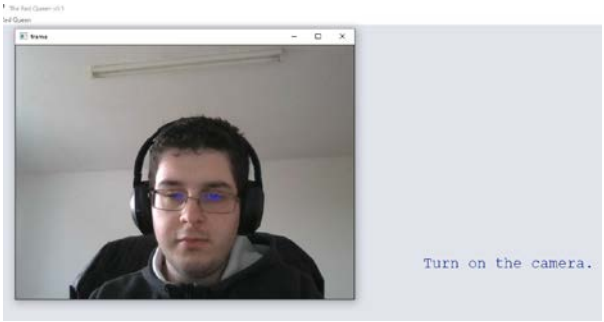
Po postopku, prikazanem na sliki 4, program zažene modul Wikipedia, poišče podatke, če podatki obstajajo, in jih uporabniku prikaže v obliki datoteke, kar vidimo na sliki 5.



Slika 5: Rezultate iskanja podatkov o osebi

Po tem se rezultati iskanja shranijo v datoteko z imenom iskalnega izraza, tako da si jih lahko uporabnik ogleda pozneje.

Primer uporabe odpiranje, kot multimedijskega ukaza, pa vidimo na sliki 6.



Slika 6: Odpiranje kamere preko Tkinter in Google Cloud Speech API (Python speech, Wit.ai Speech).

5 Zaključek

S pomočjo knjižnice Tkinter smo pripravili vmesnik do sistema za pretvorbo zvoka v besedilo, ter mu dodali možnost upravljanja enostavnih multimedijskih funkcij računalnika.

Tako smo uspešno ustvarili vmesnik do inteligentnega virtualnega asistenta z uporabo tehnologije pretvorbe zvoka v besedilo in tehnologije za razvrščanje besedil.

Čeprav je komunikacijska tehnologija znatno napredovala do stopnje, ko lahko tudi oseba z malo programskega znanja s tehnologijo nekaj proizvede, je tudi jasno, da mora znanje javnosti o tehnologiji še rasti, kar dokazuje dejstvo, da znanstveniki še vedno navajajo uporabo programskega jezika AIML kot orodja za prepoznavanje vzorcev v besedilu [12], čeprav je tehnologija naključnega učenja v zadnjih nekaj letih zelo hitro napredovala in je objektivno boljša izbira za stvari, kot je razumevanje jezika, kjer so vhodni podatki nepredvidljivi in občutljivi na majhne spremembe. Pri nadaljnjem razvoju bi lahko programu dodali več funkcij, ki bi bile bolj inteligentne in enostavnejše za uporabo. Drugi predlogi vključujejo izboljšanje videza grafičnega vmesnika, uporabo sodobnejše knjižnice za grafični vmesnik in dodajanje možnosti dela brez internetne povezave, kar pomeni uporabo tehnologij za pretvorbo govora v besedilo in klasifikacijo besedila lokalno. Kasneje bi bilo dobro, če bi dodali čustva in osebnost, kot je opisano v prispevku [10].

6 Literatura

[1] A. Zamuda, Operacijske raziskave logističnih, transportnih in ekonomskih sistemov : zbrano gradivo, 2020.

[2] S. Guaman, A. Calvopina, P. Orta, F. Tapia, S. Guun Yoo, Device Control System for a Smart Home using Voice Commands: A Practical Case, 10th International Conference2018.

[3] A.Tulshan, S. Dhage, Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa, International symposium on signal processing and intelligent recognition systems (pp. 190-201). Springer, Singapore, 2018.

[4] M. Pinola, Speech Recognition Through the Decades: How We Ended Up With Siri, PCWorld. Viewed 8 March 2021, https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html.

[5] G. McLean, Kofi Osei-Frimpong. “Hey Alexa examine the variables influencing the use of artificial intelligent in-home voice assistants.” Comput. Hum. Behav. 99, p. 3 (2019).

[6] L. Butler, “HEY GOOGLE, HELP ME LEARN” Voice Assistant Devices in the New Zealand Primary School. p.3, 2020.

[7] A. Minz, C. Mahobiya, "MR Image Classification Using Adaboost for Brain Tumor Type," *2017 IEEE 7th International Advance Computing Conference (IACC)*, Hyderabad, India, p3, 2017.

[8] A. Zamuda, C. Zarges, G. Stiglic, G. Hrovat. Stability selection using a genetic algorithm and logistic linear regression on healthcare records. Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO 2017), 143-144, 2017.

[9] C. Deuerlein, M. Langer, J. Seßner, P. Heß, J.Franke, Human-robot-interaction using cloud-based speech recognition systems, Procedia CIRP, p.1, 2021.

[10] A. M. Rahman, A. Al Mamun, A. Islam, Programming challenges of chatbot: Current and future prospective, 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE., p.1, 2017.

[11] P. Singh, B. Barry, H., Liu,, 2004. Teaching machines about everyday life. BT Technology Journal, 22(4), p1. 2004.

[12] M. S. Satu, M. H. Parvez, S. Al Mamun, Review of integrated applications with AIML based chatbot, 2015 International Conference on Computer and Information Engineering (ICCIE), Rajshahi, Bangladesh, p.3, 2015.