Natančen razpoznavalnik slovenskega govora za področje medicine

Projekt: PoVeJMo
Izdelek: D4.3
Tip izdelka: model nevronske mreže
Datum objave: september 2025

Kratek opis izdelka

V okviru projekta 4, VeMo-Med, preučujemo uporabnost govornih in jezikovnih tehnologij na področju medicine. Specifika medicinske domene, ki vpliva na uporabnost omenjenih tehnologij, je obsežna in specializirana terminologija (latinski izrazi, tujke, kratice in okrajšave), velika raznolikost govorcev ter pogosta prisotnost šuma in nestandardnih pogojev snemanja v kliničnih in ambulantnih okoljih. Obenem je zaradi občutljivosti zdravstvenih podatkov govorne in besedilne vire za učenje modelov težko pridobiti.

Izdelek D4.3 predstavlja natančen razpoznavalnik slovenskega govora (model nevronske mreže), ki je specializiran za področje medicine. Model je neodvisen od govorca, kar pomeni, da deluje zanesljivo za poljubnega govorca brez predhodne prilagoditve, in robusten na šum, kar omogoča njegovo uporabo v realnih kliničnih in ambulantnih okoljih.

Razpoznavalnik omogoča uporabo v odloženem in sprotnem načinu (angl. streaming mode) ter je dostopen prek aplikacijskega programskega vmesnika.

Razpoznavalnik za delovanje potrebuje GPU tipa NVIDIA. Končni model je lastniški in je v lasti podjetja VITASIS, saj je ta prispeval učne podatke in bazni model. Če želite razpoznavalnik uporabljati, pišite na naslov info@vitasis.si.

Razvoj in učenje modela

Razpoznavalnik je nastal v sodelovanju med konzorcijskima partnerjema UL FRI in VITASIS. VITASIS je prispeval domensko znanje s področja medicine, učne podatke in bazni model, UL FRI pa je sodelovala pri učenju modelov ter raziskala vpliv velikih jezikovnih modelov (LLM) na kakovost razpoznave govora in omejitve njihove uporabe.

Za bazni model je bil izbran model za splošno domeno arhitekture Fast Conformer Hybrid z 0,6 milijarde parametrov (0.6b), ki ga je prispeval VITASIS in je bil naučen na 20.000 urah raznolikih zvočnih posnetkov.

Bazni model je bil nato dodatno učen (finetuning) še 50 epoh na 3.000 urah generiranih oziroma sintetiziranih diktatov s področja medicine. Sintetizirani diktati so bili

uporabljeni izključno z namenom pokritja medicinske terminologije, pri sintezi pa so bili uporabljeni različni glasovi različnih ponudnikov sinteze govora, da bi zagotovili čim večjo raznolikost in s tem neodvisnost modela od govorca.

Vpliv velikih jezikovnih modelov

V skladu s cilji programa smo preučili, ali je mogoče natančnost razpoznave dodatno izboljšati z uporabo velikih jezikovnih modelov, ki s svojim obsežnim kontekstualnim znanjem potencialno pripomorejo k pravilnejši razpoznavi (na primer pri razreševanju dvoumnih izrazov, kratic in terminologije).

Vpliv velikih jezikovnih modelov na razpoznavo slovenskega govora je bil obširno in poglobljeno raziskan. V študiji so bile sistematično ovrednotene štiri skupine metod za vključevanje velikih jezikovnih modelov v sistem za razpoznavanje govora, od plitke integracije (ponovno ocenjevanje hipotez, pozivanje, preslikava hipotez v transkript) do globoke integracije (večmodalni model SALM), preizkušene z osmimi velikimi jezikovnimi modeli na šestih slovenskih evalvacijskih množicah ter ovrednotene tako po točnosti (WER) kot po hitrosti napovedovanja (RTFX). Velik jezikovni model se izkaže za najučinkovitejšega, kadar uporablja hipoteze razpoznavalnika in mu je način uporabe strogo določen oziroma je za nalogo doučen (kot ocenjevalec hipotez ali pri uglašeni preslikavi hipotez v transkript), ne pa pri prostem tvorjenju transkriptov.

Analiza računske zahtevnosti je pokazala, da so vse metode z velikimi jezikovnimi modeli občutno počasnejše in pomnilniško zahtevnejše od izhodiščnega razpoznavalnika. Raba velikih jezikovnih modelov torej lahko izboljša kakovost razpoznave, a v cevovod vnese zakasnitev, ki je praviloma prevelika za uporabo v sprotnem načinu oziroma pri nareku v realnem času. Zato veliki jezikovni modeli v predstavljenem razpoznavalniku niso vključeni v verigo sprotne razpoznave; njihova uporaba je smiselna predvsem pri odloženi obdelavi, kjer zakasnitev ni kritična. Sprotni način razpoznave temelji na računsko učinkovitih akustičnih in jezikovnih modelih, ki zagotavljajo nizko zakasnitev.

Podrobnosti raziskave so na voljo v magistrskem delu: Anton Klemen, »Uporaba velikih jezikovnih modelov za izboljšanje razpoznavanja slovenskega govora«, magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko ter Fakulteta za elektrotehniko, magistrski študijski program druge stopnje Multimedija, mentor prof. dr. Marko Bajec, Ljubljana, 2026. Izvorna koda je na voljo na naslovu https://github.com/aklemen/slollmasr.

Evaluacija

Evalvacija razpoznavalnika je bila izvedena na 100 urah dejanskih diktatov s področja medicine, ki niso bili uporabljeni pri učenju modela. Razpoznavalnik je na tej testni množici dosegel stopnjo napačno razpoznanih besed (WER) 0,0162 (1,62 %), kar potrjuje njegovo visoko natančnost na medicinski domeni.