Natančen in robusten večjezični model razpoznave govora za južnoslovanske jezike

Projekt: PoVeJMo
Izdelek: D5.4
Tip izdelka: model nevronske mreže
Datum objave: september 2025

Kratek opis izdelka

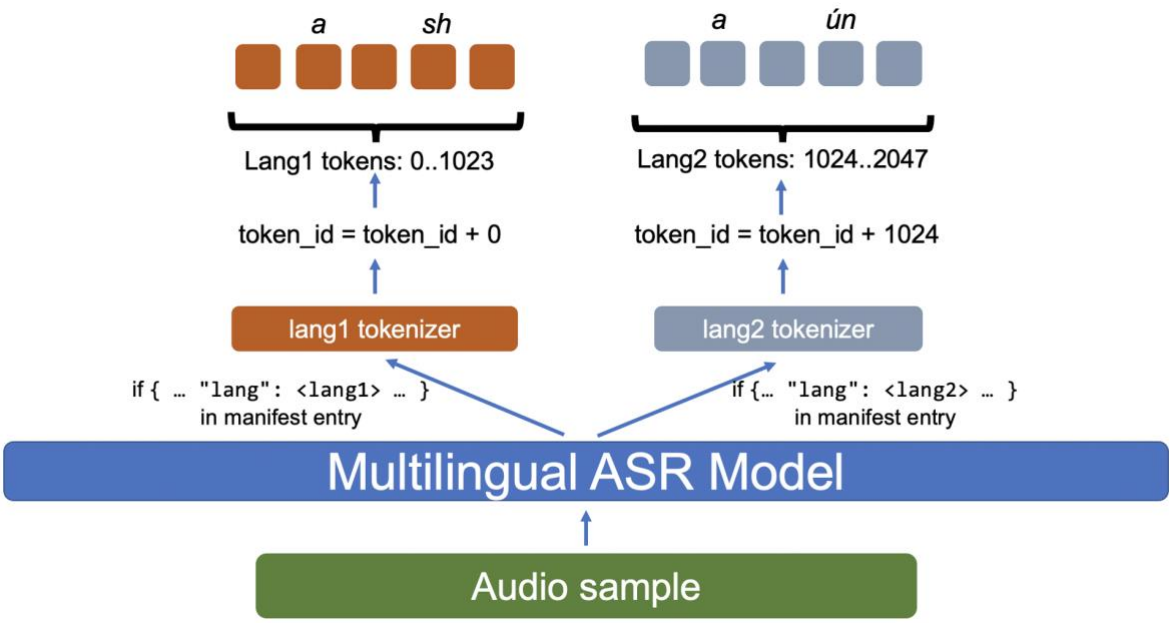
V okviru projeta 5, VeMo-IND, preučujemo uporabnost govornih in jezikovnih tehnologij v industrijskih okoljih. Specifika industrijskih okolj, ki vpliva na uporabnost omenjenih tehnologij, je prisotnost hrupa (težko zagotovimo čiste in jasne posnetke) ter uporaba različnih narečij in jezikov s strani uporabnikov.

Izdelek D5.4 predstavlja model nevronske mreže, ki omogoča prepoznavo več jezikov hkrati (z istim modelom).

Obstajata dva splošna pristopa za razvoj večjezičnih modelov. Učne podatkovne množice različnih jezikov lahko preprosto združimo v skupno učno množico in tako naučimo model. Paziti moramo pri tokenizatorju, saj mora ta podpirati vse vključene jezike. Tak pristop je enostaven, pri učenju nam ni potrebno navajati, za kateri jezik gre. To ima tudi svoje posledice; v času inference bomo težko ugotovili, četudi bo prepoznava pravilna, za kateri jezik gre.

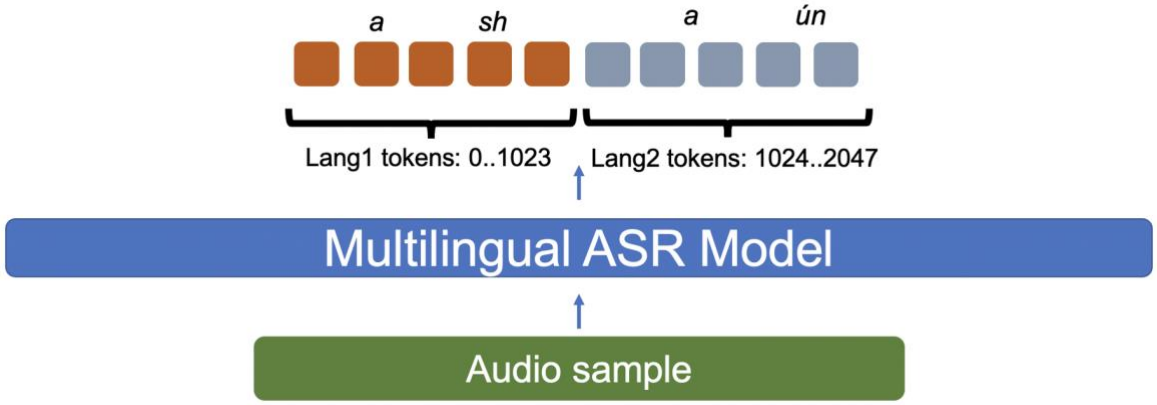
Drugi pristop temelji na ideji ponovne uporabe obstoječih že naučenih enojezičnih tokenizatorjev za vsak jezik ter njihovega enostavnega združevanja. Tokenizatorji za posamezne jezike so lahko naučeni na velikih besedilnih korpusih (za čim širšo posplošitev), izluščeni iz obstoječih enojezičnih modelov ali pa, kot običajno, naučeni na enojezičnih ASR referenčnih transkriptih.

V našem primeru smo uporabili drugi pristop. Slika 1 prikazuje način delovanja v času učenja. Za vsak jezik, ki ga želimo vključiti, moramo pripraviti učno množico ter jeziku dodeliti identifikator. Če npr. gradimo slovensko-angleški model, lahko jezik sl-SI označimo z ID=1, jezik en-US pa z ID=2. Tudi v učni množici moramo označiti, za kateri jezik gre. Dekodirnik (zadnji sloj nevronske mreže) je sestavljen iz tokenizatorjev vseh vključenih jezikov. Če so tokenizatorji veliki 1024 žetonov, bo prvih 1024 pripadalo jeziku z ID=1, drugih 1024 jeziku z ID=2 itd.



Slika 1: Sestavljanje tokenizatorjev v času učenja (vir slike)

V času učenja bo model za vsak žeton vrnil verjetnostno distribucijo čez celoten sestavljen tokenizator. Če bo npr. imel žeton na poziciji 2000 največjo verjetnost, bomo vedeli, da pripada jeziku ID=2, saj žetoni tega jezika na pozicijah od 1025 do 2048.



Slika 2: Predikcija z uporabo več-jezičnega modela (vir slike)

Več-jezični model sl, hr, sr, en

V okviru projekta PoVeJMo smo izdelali več-jezični model, ki vključuje naslednje jezike:

1. Slovenščina (15000 ur, učna množica konzorcijskega partnerja VITASIS)
2. Hrvaščina (4000 ur, učna množica konzorcijskega partnerja VITASIS)
3. Srbščina (2000 ur, CLARIN ParlaMint)
4. Angleščina (10000 ur, učna množica konzorcijskega partnerja VITASIS)

Angleščina je bila dodana po pogvoru s konzorcijskim partnerjem ŠPICA in njegovim poznavanjem potreb v zvezi z več-jezičnostjo.