Razpoznavanje tekočega govora v slovenščini z bazo predavanj SI TEDx-UM

Andrej Žgank, Darinka Verdonik, Mirjam Sepesy Maučec

Inštitut za elektroniko in telekomunikacije
Fakulteta za elektrotehniko, računalništvo in informatiko
Univerza v Mariboru
Smetanova ul. 17, 2000 Maribor
andrej.zgank@um.si, darinka.verdonik@um.si, mirjam.sepesy@um.si

Povzetek

V članku bomo predstavili novi slovenski govorni vir, nastal na osnovi posnetkov predavanj TEDx. Govorna baza vsebuje posnetke 242 predavanj, v skupni dolžini 54 ur. Transkribiranje govora v bazi SI TEDx-UM smo izvedli v dveh delih. Učni nabor smo transkribirali avtomatsko, z uporabo razpoznavalnika govora UMB Broadcast News. Razvojni in testni nabor baze, ki obsega 3 ure govornega materiala, pa smo transkribirali ročno, v skladu z nadgrajenimi priporočili za transkribiranje korpusa GOS. Razpoznavalnik govora UMB Broadcast News ASR je prvenstveno namenjen televizijskim oddajam, zato smo v nadaljevanju izvedli analizo, kakšen vpliv ima tip jezikovnega modela na uspešnost razpoznavanja govora v bazi SI TEDx-UM. Primerjali smo privzeti jezikovni model domene televizijskih oddaj s splošnim jezikovnim modelom grajenim na korpusu FidaPLUS. Povprečna napaka razpoznavanja besed na testnem naboru baze SI TEDx-UM je znašala 50,7%. Govorna baza SI TEDx-UM je prosto dostopna.

Slovenian Continuous Speech Recognition with the SI TEDx-UM Talks Database

This paper presents a new Slovenian spoken language resource for automatic speech recognition. The SI TEDx-UM speech database contains 242 different Slovenian TEDx talks in total length of 54 hours. The training set was transcribed automatically using the UMB Broadcast News speech recognition system. The development and evaluation set were manually transcribed, using adapted transcription guidelines for the Slovenian GOS corpus. The influence of language model domain was also analysed. The UMB Broadcast News ASR language model was compared with the general language model build on FidaPLUS corpus. The average word error rate of 50.7% was achieved on the SI TEDx-UM evaluation set. The SI TEDx-UM speech database is freely available.

1. Uvod

Avtomatsko razpoznavanje tekočega govora predstavlja eno izmed pomembnih IKT tehnologij, tako na področju vmesnikov človek-stroj kot tudi na področju digitalnih vsebin. Porast le-teh v zadnjih nekaj letih ima za posledico nujnost avtomatske obdelave velike množice takšnega medijskega materiala. Razpoložljivost ustreznih govornih virov predstavlja še vedno eno izmed kritičnih točk na področju razvoja govornih tehnologij, predvsem avtomatskega razpoznavanja govora. Klasičen pristop izdelave govornih virov je zaradi ročnega transkribiranja dolgotrajen in drag. Posledično vlada za velik delež svetovnih jezikov, med katerimi je tudi slovenščina, pomanjkanje ustreznih govornih virov.

V članku[1] bomo predstavili nov slovenski jezikovni vir, govorno bazo SI TEDx-UM, ki je nastala na osnovi sklopa slovenskih predavanj TEDx. Ta predavanja so v svetu uveljavljena že vrsto let, v zadnjih petih letih so se širše uveljavila tudi v Sloveniji. Predavanja pokrivajo različne aktualne tematike s področja tehnologije, izobraževanja, umetnosti in družbe.

Govorna baza SI TEDx-UM je razdeljena na dva dela. Prvi del predstavlja učni nabor, za katerega smo segmentacijo in transkripcije pripravili avtomatsko, z uporabo razpoznavalnika slovenskega govora. Na takšen način smo uspeli bistveno poenostaviti in pohitriti izdelavo govorne baze, vendar bo potrebno zaradi prisotnih napak v transkripciji v prihodnje uporabljati drugačne pristope razvoja razpoznavalnika govora (Rousseau et al., 2014). Drugi del govorne baze predstavlja razvojni in testni nabor, kjer smo transkripcije predavanj tvorili ročno. Za izdelavo izhodiščne verzije avtomatskih transkripcij smo zaradi njegovega robustnega delovanja uporabili obstoječi avtomatski razpoznavalnik govora UMB Broadcast News (Žgank et al., 2014a), kjer so modeli specifično uravnotežen na domeno televizijskih informativnih oddaj. V prispevku bomo analizirali, kako razlika med splošnim jezikovnim modelom in takšnim, uravnoteženim na televizijske informativne oddaje, vpliva na uspešnost razpoznavanja govora predavanj TEDx.

V nadaljevanju članka bomo najprej predstavili postopek zajema in ročnega transkribiranja govorne baze SI TEDx-UM. V tretjem poglavju bomo predstavili postopek avtomatske segmentacije in transkribiranja zajetega govornega materiala. V četrtem poglavju bo sledila predstavitev dveh tipov jezikovnih modelov, uporabljenih za avtomatsko transkribiranje. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v petem poglavju. Zaključek in smernice za nadaljnje delo bomo podali v šestem poglavju.

2. Slovenska govorna baza SI TEDx-UM

2.1. Zajem materiala

Izvorni material za govorno bazo smo zajeli s spletne strani YouTube, kjer so na voljo posnetki različnih predavanj TEDx v slovenskem jeziku. Ker so predavanja prirejali različni organizatorji, smo na takšen način najlažje zajeli vsa dostopna slovenska predavanja, ki jih je

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069.

bilo več kot 300. Posnetki predavanj so na spletni strani YouTube na voljo v različnih izgubnih kodekih (tipično: zvok: MPEG AAC, video: H.264). Vedno smo uporabili tistega, ki zagotavlja najvišjo možno kakovost govornega materiala. Izvorne govorne posnetke smo pretvorili v format WAV s frekvenco vzorčenja 16 kHz in 16-bitno ločljivostjo, kar je standardni format govorne baze BNSI Broadcast News in kot takšen zadostuje za razvoj razpoznavalnikov govora. Kakovost video materiala je bila pri zajemu drugotnega pomena, saj je video služil zgolj kot pomoč pri transkribiranju govora.

2.2. Transkribiranje

Transkribiranje spontanega govora v jezikovnih virih se je v slovenščini v preteklih letih razvijalo predvsem ob posameznih govornih bazah, kot sta bili BNSI Broadcast News (Žgank et al., 2004) in SiBN Broadcast News (Žibert in Mihelič, 2004), pomembno prelomnico pa predstavlja standard transkribiranja, razvit ob referenčnem govornem korpusu GOS (Verdonik et al., 2013). Novosti v načelih transkribiranja, ki jih je prinesel GOS v primerjavi z bazo BNSI Broadcast News, so v preteklosti že bile podrobno analizirane (Žgank et al., 2014). Pri snovanju specifikacij transkribiranja za slovensko bazo TEDx smo prepoznane razlike v glavnem upoštevali ter pripravili posodobljena načela transkribiranja, ki upoštevajo standarde, vzpostavljene s korpusom GOS in se v primerjavi z bazo BNSI Broadcast News najbolj razlikujejo v načinu segmentiranja izjav na osnovne enote – segmente oz. izjave – in v načinu zapisovanja govora, v primerjavi s korpusom GOS pa vključujejo bolj natančno označevanje akustičnega ozadja in akustičnih dogodkov.

Tako so v SI TEDx-UM osnovne enote segmentirane po načelu, da segmenti približno ustrezajo pojmu izjave, pri čemer izjavo razumemo kot osnovno enoto govora, ki približno ustreza pojmu (kratke) povedi v pisni rabi. Iz tehničnih razlogov smo pri tem upoštevali omejitev, da morajo vsak segment vedno zamejevati vsaj tolikšni premori v govoru, da je mogoče postaviti časovno mejo med segmentoma na način, da ni v zvočnem signalu odrezan noben delček fonema predhodne ali naslednje besede. Prednost imajo vedno krajši segmenti, saj to omogoča uspešnejše učenje akustičnih modelov.

Zapis govora namesto enojnega standardiziranega zapisa vključuje dvotirni sistem zapisovanja, pogovorni in standardizirani, uveden s korpusom GOS. Posebne odločitve so bile potrebne le v zvezi s predhodno ugotovljenimi pomanjkljivostmi zapisovanja v korpusu GOS (Verdonik 2014). Pri ugotovljenih nedoslednostih zapisovanja zvočnika dvoustnični v (ni nosilec zloga), kjer v korpusu GOS najdemo po večkrat tudi pogovorne zapise tipa *mau* (*malo*), *biu* (*bil*), *šou* (*šel*), *dou* (*dol*), *prou* (*prav*), *dau* (*da bo*), *nou* (*ne bo*) itd., ohranimo pravilo, da ga zapisujemo s črko »v« (*prov, nav, navm, odpravt, davn...*) oz. tudi z »l«, če tako izhaja iz knjižne norme (*kosil, mel*), in smo še posebej pozorni na ugotovljene nedoslednosti. Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog *v*, izgovorjen samoglasniško), pa ga enako kot prej pišemo s črko »u« (*pršu, vidu, u tem delu...*). Pri zapisovanju neverbalnih in polverbalnih glasov se držimo seznama, predstavljenega v (Verdonik 2014). Določni člen 'ta' v tipu 'ta rdeči' po novem pišemo skupaj v pogovornem zapisu, narazen pa v standardiziranem zapisu.

Pri označevanju akustičnih dogodkov in ozadja je novost dodatno označevanje akustičnega ozadja, ki traja 3 sek. ali več in se spremeni v primerjavi s tem, kakšno ozadje prevladuje v posnetku večino časa. Za to se uporablja posebna časovna sled. Za seznam akustičnih dogodkov smo uporabljali oznake, podane znotraj orodja Transcriber AG (vključno s funkcijo named entities), je pa bolj natančno in označeno vedno, ko je kak akustični dogodek slišen, ne samo takrat, ko je pragmatično pomemben za komunikacijo.

Transkribiranje je potekalo s pomočjo prenovljene različice programa Transcriber, Transcriber AG. Čeprav orodje ponuja med drugim izredno dobrodošlo novost, da omogoča transkribiranje ob videoposnetku, se je med delom žal pokazalo kot izredno nestabilno ter z nekaj napakami v delovanju zlasti pri označevanju akustičnih dogodkov.

2.3. Podatki o bazi

V končno verzijo govorne baze smo vključili ročno izbranih 242 predavanj, katerih značilnosti so ustrezale kriterijem na področju razpoznavanja govora. Glavna značilnost izločenih predavanj je bila popačena akustična karakteristika kot posledica različnih vzrokov: prekrivanje govora več govorcev, glasbena spremljava oziroma glasbeno ozadje, nizka kakovost posnetkov ipd. Prav tako smo izločili vsa predavanja v tujem jeziku. Posledično vsebuje tipično predavanje govor enega slovenskega govorca v dobrih akustičnih pogojih. V bazi je 66% moških govorcev in 34% ženskih govork. Skupna dolžina vključenih predavanj je 54 ur, izvirajo pa iz časovnega obdobja 6 let. Izmed 242 predavanj smo izbrali 13 predavanj v skupni dolžini 3 ur, ki predstavljajo razvojni in testni nabor govorne baze. Za ta del govorne baze smo v skladu z zastavljenimi pravili izvedli ročno transkribiranje, saj lahko samo na takšen način govorno bazo uporabljamo za eksperimente razpoznavanja tekočega govora. Transkripcije za učni nabor baze v dolžini 51 ur smo pripravili avtomatsko s pomočjo razpoznavalnika slovenskega govora in na takšen način bistveno pohitrili izdelavo nove govorne baze. Transkripcije govorne baze SI TEDx-UM vsebujejo 372k pojavnic, od tega jih je 32k različnih.

3. Avtomatsko transkribiranje

Za pripravo avtomatsko tvorjenih transkripcij predavanj TEDx smo uporabili razpoznavalnik tekočega slovenskega govora UMB Broadcast News (Žgank et al., 2014a). Takšen pristop bistveno poenostavi izdelavo transkripcij, kadar lahko v njih toleriramo večji ali manjši delež napak. Prvi korak avtomatske obdelave govorne baze SI TEDx-UM je predstavljala akustična segmentacija s pomočjo modelov GMM, s katerimi smo tvorili akustično homogene dele govornih posnetkov, primerne za avtomatsko razpoznavanje govora.

Uporabljeni razpoznavalnik govora je bil naučen na kombinaciji govorne baze BNSI Broadcast News in interne govorne baze IETK-TV. Baza BNSI Broadcast News obsega 36 ur obdelanih podatkov, od tega je 30 ur namenjenih učenju akustičnih modelov, 3 ure razvojnemu prilagajanju sistema in 3 ure testiranju razpoznavalnika. Baza je bila zajeta v letih 2003 in 2004 v sodelovanju z RTV Slovenija ter vključuje informativne oddaje tega TV-programa, predvsem TV Dnevnik in Odmeve. Posnetki so bili ročno segmentirani in transkribirani s pomočjo orodja Transcriber, nato pa vključeni v razvoj razpoznavalnika govora. Z uporabo govorne baze IETK-TV smo pridobili dodatnih 29 ur transkribiranega materiala, tako da je končni učni nabor razpoznavalnika govora za transkribiranje baze SI TEDx-UM zajemal 59 ur posnetkov.

Razpoznavalnik govora UMB Broadcast News uporablja za izločanje značilk 12 mel kepstralnih koeficientov z energijo ter prvimi in drugimi odvodi, tako da ima končni vektor 39 elementov. Dodali smo še normalizacijo srednjih vrednosti kepstralnih koeficientov, saj na tak način zmanjšamo razlike med različnimi snemalnimi okolji, kar je izredno pomembno v primeru kombiniranja več govornih baz. Akustični modeli razpoznavalnika govora so bili naučeni z iterativnim postopkom in so v končni obliki vsebovali kontekstno odvisne grafemske medbesedne modele s kombinacijo 32 zveznih Gaussovih porazdelitvenih funkcij verjetnosti na stanje. Več podrobnosti o uporabljenem sistemu UMB Broadcast News je podanih v (Žgank et al., 2014a). V drugem koraku avtomatske obdelave govorne baze SI TEDx-UM smo izvedli razpoznavanje govora na akustično homogenih posnetkih, ki so bili rezultat prvega koraka.

4. Jezikovna modela

Za gradnjo jezikovnih modelov smo uporabili orodje SRI Language Modeling Toolkit (Stolcke, 2002). Izhodiščni jezikovni model, ki smo ga zgradili v okviru razpoznavalnika UMB Broadcast News, je interpolirani 3-gramski model, sestavljen iz štirih komponent. Pri učenju vseh štirih komponent smo uporabili Good-Turingovo glajenje in sestopanje po Katzu. Glede na velikost učnega korpusa in glede na utež v končnem modelu je največja komponenta FidaPLUS. Le-ta je grajena na korpusu FidaPLUS, ki predstavlja referenčno zbirko vsakdanje javne rabe slovenščine v pisnih besedil v obdobju med 1990 do 2006, in vsebuje 621 milijonov besed (Arhar in Gorjanc, 2007). Ostale tri komponente (BNSI-Speech, BNSI-Text in Večer) imajo približno enako utež in torej v enakem deležu prispevajo h končni oceni verjetnosti jezikovnega modela. Tudi komponenta Večer je predstavnik pisanega jezika, ostali dve pa govorjenega. Interpolacijske uteži so bile določene na razvojni množici BNSI-Devel, ki je po strukturi enaka BNSI-Speech in torej predstavlja reprezentativni vzorec ciljne domene za razpoznavalnik UMB Broadcast News. Pričakovano je, da je uspešnost jezikovnega modela BNSI na bazi TEDx-UM, ki predstavlja prehod na novo domeno, manjša.

Slovar razpoznavalnika govora obsega 64.000 besed in je prilagojen domeni BNSI, saj je v pretežni meri sestavljen iz besed korpusov BNSI-Speech in BNSI-Text. Za novo domeno bi bilo sicer smiselno sestaviti nov slovar, a nas v članku zanima, kakšno pokritost nove domene daje obstoječi jezikovni model. Po drugi strani pa nove besede v slovarju razpoznavanja ne izboljšajo, če jezikovni model nima znanja o pojavitvah teh besed. Res je, da na ta način zmanjšamo OOV, ni pa nujno, da izboljšamo razpoznavanje, saj majhne verjetnosti jezikovnega modela zelo verjetno hipoteze, ki vsebujejo nove besede, izločijo iz nabora najverjetnejših hipotez.

Predavanje	Tematika	JM1 PP	JM2 PP	OOV
1	potovanja	409	431	21%
2	tehnologija	390	412	23%
3	družba	440	475	22%
4	tehnologija	379	405	28%
5	umetnost	481	506	26%
6	družba	491	491	26%
7	znanost	323	336	22%
8	znanost	242	234	20%
9	družba	429	451	27%
10	umetnost	400	399	24%
11	družba	428	451	19%
12	znanost	402	412	24%
13	družba	287	260	23%
vsa	različna	390	403	24%
BNSI eval	različna	247	387	4%

Tabela 1: Rezultati jezikovnih modelov UMB Broadcast News in FidaPLUS na testnih vzorcih SI TEDx-UM.

Glede na to, da sta si domeni BNSI in TEDx precej različni, smo preverili tudi uspešnost jezikovnega modela, učenega samo na korpusu FidaPLUS, ki predstavlja bolj splošno domeno. Korpus FidaPLUS je referenčna zbirka pisnih besedil in s tem, ko iz jezikovnega modela odstranimo komponente, ki modelirajo govorjeno rabo, izgubimo modelirane odvisnosti, tipične za govorjeno rabo, ki jih v pisni rabi ni ali pa so zelo redke in se v velikem korpusu zameglijo. Opozoriti velja tudi na značilnosti spontanega govora, ki jih niti jezikovni model BNSI ne modelira, saj gre pretežno za brani govor.

5. Rezultati

Izvedli smo eksperimentalno primerjavo rezultatov razpoznavanja govora z uporabo dveh različnih jezikovnih modelov: interpoliranega na besedilih iz televizijskih informativnih oddaj ter splošnega, grajenega izključno na besedilnem korpusu FidaPLUS. Najprej smo primerjali uspešnost obeh jezikovnih modelov na testnih vzorcih SI TEDx-UM. Rezultati so zbrani v tabeli 1. JM1 PP je perpleksnost jezikovnega modela UMB Broadcast News, JM2 PP pa jezikovnega modela FidaPLUS. Ker oba jezikovna modela uporabljata isti slovar, je delež OOV za oba enak. Perpleksnosti jezikovnega modela FidaPLUS so

boljše le na vzorcih 8 in 13, na vseh ostalih pa slabše. Sklepamo, da je modeliranje govorjene rabe jezika pomembna komponenta jezikovnega modela. Visok delež besed izven slovarja kaže na različnost domen BNSI in TEDx, velike vrednosti perpleksnosti obeh jezikovnih modelov pa na specifičnost domene TEDx.

V drugem koraku vrednotenja obeh jezikovnih modelov smo izvedli eksperimente razpoznavanja govora na ročno transkribiranem testnem naboru baze SI TEDx-UM. Rezultati so podani v tabeli 2 v obliki napake razpoznanih besed (NRB).

Predavanje	JM1 NRB(%)	JM2 NRB(%)
1	50,5	51,3
2	54,7	56,6
3	57,7	58,5
4	39,2	38,5
5	67,1	67,6
6	46,1	45,3
7	52,9	53,3
8	35,5	34,9
9	51,4	52,9
10	35,0	35,5
11	52,4	51,0
12	70,3	69,3
13	38,9	35,1
vsa	50,7	50,7
BNSI eval	26,6	26,6

Tabela 2: Napaka razpoznavanja govora na testnih vzorcih baze SI TEDx-UM za oba jezikovna modela.

Oba jezikovna modela sta dosegla 50,7% napako razpoznavanja besed, kar kaže na to, da imata specifični jezik in tematika predavanj bistveno večjo težo v primerjavi z različnimi besedilnimi viri, ki smo jih vključili v jezikovni model. Analiza na nivoju posameznih predavanj je pokazala, da je NRB podobna, ne glede na uporabljeni jezikovni model. Do edinega odstopanja je prišlo v primeru predavanja številka 13, kjer je JM2 dosegel za 3,8% boljši rezultat. Najboljši rezultat za posamezno predavanje je napaka razpoznavanja besed 34,9%, dosežena pri predavanju številka 8, ki ima tudi najnižjo perpleksnost jezikovnega modela.

Primerjava z bazo BNSI je pokazala da so doseženi rezultati razpoznavanja govora slabši kot tisti, doseženi na govorni bazi BNSI Broadcast News, kar lahko pripišemo bistveno višjemu deležu besed izven slovarja (OOV) ter razlikam v domeni in karakteristikah med obema govornima viroma.

6. Zaključek

Govorna baza SI TEDx-UM predstavlja pomemben nov vir za razvoj govornih tehnologij, saj odpira nove možne tematike raziskovalnega dela za slovenski jezik. Z uporabo takšnih pristopov bo možno v prihodnje bistveno učinkoviteje graditi nove govorne vire, ki so še vedno neobhodno potrebni za nadaljnji razvoj govornih tehnologij, kot je avtomatsko razpoznavanje tekočega govora. Prvi rezultati razpoznavanja govora na bazi SI-TEDx-UM kažejo velik vpliv tematike predavanj, ki pomembno odstopa od trenutno obstoječega sistema za razpoznavanje govora.

V prihodnje bomo skušali rezultat razpoznavanja govora izboljšati s prilagojenimi jezikovnimi modeli, ki jih bomo gradili na novih jezikovnih virih, predvsem na transkribiranem gradivu TEDx in gradivu TED v drugih jezikih, ki je prevedeno v slovenščino.

Govorna baza SI TEDx-UM je v skladu z licenco Creative Commons 3.0 prosto dostopna na spletni strani Inštituta za elektroniko in telekomunikacije UM FERI: http://ietk.feri.um.si/en/portfolio/sitedxumenglish.

7. Literatura

Špela Arhar in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2, 95--110.

Anthony Rousseau et al.. 2014. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. *Proc. of the LREC'14,* Reykjavik, Islandija.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. *International Conference on Speech and Language Processing*, II: 901--904.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language resources and evaluation*, 47(4):1031-1048.

Darinka Verdonik. 2014. Vprašanja zapisovanja govora v govornem korpusu Gos. *Jezikovne tehnologije. IS 2014*, Ljubljana, Slovenija, str. 151-156.

Andrej Žgank, Tomaž Rotovnik, Darinka Verdonik, Zdravko Kačič. 2004. Baza broadcast news za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. *Jezikovne tehnologije. IS 2004,* Ljubljana, Slovenija, str. 98-97.

Andrej Žgank, Ana Zwitter Vitez, Darinka Verdonik. 2014. The Slovene BNSI broadcast news database and reference speech corpus GOS: towards the uniform guidelines for future work. *Proc. of the LREC'14*, Reykjavik, Islandija, str. 2644-2647.

Andrej Žgank, Gregor Donaj, Mirjam S. Maučec. 2014a. Razpoznavalnik tekočega govora UMB Broadcast News 2014 : kakšno vlogo igra velikost učnih virov? *Jezikovne tehnologije. IS 2014,* Ljubljana, Slovenija, str.147-150.

Janez Žibert, France Mihelič. 2004. Development, evaluation and automatic segmentation of Slovenian broadcast news speech database. *Jezikovne tehnologije. IS 2004*, Ljubljana, Slovenija, str. 72-78.