Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj in Zdravko Kačič

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova ul. 17, SI-2000 Maribor, Slovenija
andrej.zgank@uni-mb.si http://www.dsplab.uni-mb.si

Povzetek

V članku bomo predstavili nov slovenski jezikovni vir – bazo SloParl. Sestavlja jo material zajet v okviru parlamentarnih razprav v Državnem zboru Republike Slovenije. Glavno vodilo projekta je bilo na stroškovno učinkovit način izdelati nov slovenski jezikovni vir za avtomatsko razpoznavanje tekočega govora. V govornem korpusu SloParl se tako nahaja skupno 100 ur govornega materiala. Le-ta je namenjen nenadzorovanemu oziroma rahlo nadzorovanemu učenju akustičnih modelov. V skladu s tem je potekala tudi priprava transkripcij govornega materiala. Drugi del baze SloParl tvori tekstovni korpus, ki vsebuje obdelane magnetograme parlamentarnih razprav iz obdobja 1996 - 2005. Tekstovni korpus vsebuje 23M besed. Primerjava z obstoječimi slovenskimi jezikovnimi viri je pokazala, da baza SloParl uspešno pokriva nove vidike na področju modeliranja slovenskega govora.

Slovenian parliamentary debates speech and text database for automatic speech recognition

A novel Slovenian language resource – the SloParl database – will be presented in this paper. It consists from spoken material acquired in the Slovenian Parliament. The main goal of the project was to cost-effectively collect a new Slovenian language resource that could be used to augment the available Slovenian speech corpora for developing a large vocabulary continuous speech recognition system. The SloParl speech corpus has a total length of 100 hours. The SloParl speech corpus will be used for lightly supervised or unsupervised acoustic models' training. In accordance with this, the accompanying transcriptions were prepared. The second part of the SloParl database is the text corpus, which covers text of all debates between years 1996 - 2005. It consists from 23M words. Comparison with other Slovenian language resources showed that SloParl database adds new aspects to the modelling of Slovenian language.

1. Uvod

Področje avtomatskega razpoznavanja govora je neločljivo povezano z izdelavo jezikovnih virov, ki so potrebni za izdelavo modelov vključenih v razpoznavalnik govora. Osnovne tri vrste jezikovnih virov so:

- *Govorna baza*: v obliki transkribiranega govora je potrebna za učenje akustičnih modelov razpoznavalnika govora.
- *Tekstovni korpus*: se uporablja za izdelavo jezikovnih modelov pri razpoznavanju tekočega govora.
- *Fonetični slovar*: služi za povezavo med obema zgornjima jezikovnima viroma.

Avtomatsko razpoznavanje govora dosega najboljše rezultate takrat, kadar so jezikovni viri, ki jih uporabimo za učenje modelov, kar najbolj podobni govornemu materialu s katerim se bo razpoznavalnik govora dejansko srečal. Posledica te zahteve je, da je velikokrat potrebno za novo področje uporabe razpoznavalnika govora pripraviti nove jezikovne vire.

Postopek izdelava jezikovnih virov za avtomatsko razpoznavanje govora je običajno zelo drag in dolgotrajen, saj je pri izdelavi kvalitetne govorne baze potrebno veliko ročnega dela. To se odraža v številu jezikovnih virov, ki so dostopni za posamezni jezik. Tukaj prevladujejo predvsem jeziki z velikim številom govorcev, za katere je izražen močan ekonomski interes za razvoj razpoznavalnikov govora.

S stališča kompleksnosti razpoznavalnikov govora so najobsežnejši jezikovni viri potrebni za razpoznavalnike tekočega govora. Če želimo pravilno oceniti parametre akustičnih in jezikovnih modelov, potrebujemo za izpeljavo postopka učenja velike količine učnih podatkov. V zadnjem obdobju lahko opazujemo razvoj razpoznavalnikov tekočega govora tudi za slovanske jezike (Byrne et al., 1999; Nouza et al., 2004; Žgank et al., 2001), kamor sodi tudi slovenščina. Le-ta je zaradi svojih značilnosti še posebej zahtevna za razpoznavanje tekočega govora.

Na področju razpoznavanja tekočega slovenskega govora je bila prva dostopna govorna baza SNABI (Dreo, 1995). Njena slabost je, da je omejena na posamezne domene. Razvoj slovenskega razpoznavalnika govora za neomejeno domeno omogoča govorna in tekstovna baza BNSI Broadcast News (Žgank et al., 2004) in govorna baza SiBN Broadcast News (Žibert in Mihelič, 2004). Kot posebni dodatek k bazi BNSI Broadcast News je bila razvita govorna baza SINOD, ki pokriva slovenski govor tujih govorcev (Žgank et al., 2006). Če primerjamo obseg obstoječih govornih baz za slovenski jezik s količino transkribiranega govornega materiala za druge jezike lahko vidimo, da slovenščina na tem področju razvoja jezikovnih tehnologij še vedno zaostaja.

Predstavljene težave in omejitve so bile vzpodbuda za razvoj novega slovenskega jezikovnega vira, baze SloParl[1]. Predstavljeni jezikovni vir vsebuje parlamentarne razprave iz Državnega zbora Republike Slovenije in je sestavljen iz

[1]Delo je bilo delno financirano s strani Agencije za raziskovalno dejavnost Republike Slovenije po pogodbi št. P2-0069.

govorne baze s transkripcijami in iz tekstovnega korpusa.

Projekt izdelave jezikovnega vira SloParl se je začel konec leta 2005 v sodelovanju med Fakulteto za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in Državnim zborom Republike Slovenije. Osnovno vodilo pri izdelavi baze SloParl je bilo v kratkem času ter z minimalnim ročnim delom zagotoviti nov slovenski jezikovni vir za razpoznavanje tekočega govora. V primeru govorne baze SloParl smo tako namesto namenskih ročno tvorjenih transkripcij, ki natanko pokrivajo vse vidike govornega signala (govor, mašila, zvoke iz ozadja, šum, lastnosti govorca,...) uporabili magnetograme, ki so nastali v državnem zboru in vsebujejo samo prepis govora. Več avtorjev (Kemp in Waibel, 1999; Wessel in Ney, 2001; Lamel et al., 2002) je takšne govorne baze uspešno uporabilo za nenadzorovan oziroma rahlo nadzorovan postopek učenja akustičnih modelov. Takšen pristop je možno učinkovito uporabiti tudi z netranskribiranim govornim materialom. Dodatna motivacija za razvoj baze SloParl je bilo dejstvo, da se baze parlamentarnih razprav (Gollan et al., 2005; Biatov in Köhler, 2002) v zadnjem času pogosto uporabljajo za razvoj najsodobnejših razpoznavalnikov tekočega govora, kot so tisti vključeni v prevajalnike govora v govor ali v multimodalne aplikacije za zajem informacij.

V nadaljevanju članka bomo v drugem poglavju opisali postopek priprave baze SloParl ter na kakšen način je bil zajet material. Lastnosti in statistiko govornega korpusa bomo predstavili v tretjem poglavju, tekstovnega pa v četrtem. Zaključek bomo podali v petem poglavju.

2. Priprava baze SloParl in zajemanje materiala

Zasedanja Državnega zbora Republike Slovenije lahko razdelimo na dva dela. V prvem delu so redne seje, ki so praviloma na sporedu enkrat na mesec in trajajo več dni. Teme sej so običajno različne, pri tem pa lahko pokrivajo zelo širok nabor vsebin. V drugi del delovanja državnega zbora sodijo izredne seje, ki so sklicane v primeru obravnave kakšne nujne teme. Izredne seje praviloma pokrivajo ozko tematiko, vezano na posamično zadevo. Predstavljena razlika med obema vrstama sej državnega zbora je pomembna pri izdelavi jezikovnega vira.

Državni zbor sestavlja 90 poslank in poslancev, ki so izvoljeni za obdobje štirih let. Sejo vodi predsednik ali podpredsednik zbora. Hkrati s poslanci sodelujejo na sejah kot govorci tudi različni drugi ljudje. Praviloma gre tukaj za člane Vlade Republike Slovenije, ki odgovarjajo na poslanska vprašanja in pobude oziroma predstavljajo predloge zakonov.

Vsaka seja državnega zbora je hkrati snemana na dva načina. Zvok in slika sta v digitalni obliki zajeta na DVD medij, hkrati pa je samo zvok v analogni obliki posnet tudi na profesionalni magnetofonski trak. Pri izbiri medija za presnemavanje govornega materiala za bazo izgovorjav je bilo prvo vodilo kvaliteta zajetega materiala. Ker v državnem zboru pri snemanju na DVD medije uporabljajo kodek z izgubno kompresijo, smo za izvor govornega materiala v bazi SloParl izbrali analogne magnetofonske trakove. Analogni zvočni signal smo zajemali neposredno na osebni računalnik v katerega je bila vgrajena zvočna kartica visoke ločljivosti SoundBlaster Audigy. Izvorni analogni govorni signal smo digitalizirali s 16 bitno ločljivostjo pri 48 kHz vzorčenju. Kasneje smo zajeti govor pretvorili v signal s 16 bitno ločljivostjo in 16 kHz vzorčenjem.

Med zajemanjem govornega signala so se občasno pojavile težave z nastavitvijo glasnosti, saj imajo različni govorci različen stil govora. Praviloma je bil med glasnejšimi govorci predsedujoči, ki smo ga tako uporabili za nastavitev referenčne vrednosti. Akustično okolje sej državnega zbora je s stališča avtomatskega razpoznavalnika govora zelo kompleksno, saj se pogosto pojavljata šum ter govor v ozadju. Velika parlamentarna dvorana lahko v določenih primerih generira odmev, kar dodatno oteži akustično okolje. Ozvočenje govorcev na sejah državnega zbora je izvedeno s konferenčnimi mikrofoni, ki so nameščeni na mizah pred parlamentarci.

V govorni del baze SloParl smo izbrali posnetke iz obdobja 2000 - 2005. Pri izbiri posnetkov parlamentarnih sej smo prednostno izbrali tiste seje, katerih datumi se prekrivajo s televizijskimi oddajami vključenimi v slovensko bazo BNSI Broadcast News (Žgank et al., 2004). Takšno prekrivanje govornega materiala je še posebej pomembno, kadar želimo zožiti domeno akustičnih in jezikovnih modelov avtomatskega razpoznavalnika govora.

Drugi del baze SloParl sestavlja tekstovni korpus. Magnetogrami parlamentarnih razprav so prosto dostopni na domači strani Državnega zbora RS za obdobje od leta 1996 naprej. Uporabljen je HTML format datotek. Magnetogrami sej vsebujejo zapise razprav, ki so jih pripravili v državnem zboru. Pripravljeni tekstovni korpus bomo uporabili predvsem za učenje jezikovnih modelov avtomatskega razpoznavalnika tekočega govora.

3. Govorni korpus

Govorni korpus slovenske baze SloParl vsebuje posnetke 100 ur govornega materiala. Celotno govorno bazo SloParl smo razdelili na tri nabore: učnega, razvojnega in testnega. Največji izmed teh treh je učni nabor, ki obsega 92 ur govornega materiala. V razvojnem naboru (redna seja državnega zbora iz junija 2001), ki je namenjen nastavljanju parametrov razpoznavalnika tekočega govora, se nahajajo 4 ure posnetkov. Preostale 4 ure govornega materiala (redna seja iz februarja 2002) smo dodelili v testni nabor, ki je namenjen vrednotenju razpoznavalnika govora. Osnovna statistika govornega korpusa baze SloParl je predstavljena v tabeli 1.

Parameter	Vrednost
Število sej	20
Redne seje	13
Izredne seje	7
Povprečna dolžina (učni set)	5:05

Tabela 1: Statistika govornega korpus slovenske baze SloParl.

Govorni korpus sestavlja 20 sej državnega zbora, od tega 13 rednih ter 7 izrednih. Povprečna dolžina seje, ki

je vključena v učni nabor, znaša 5:05 ure. Običajno vsebuje vsaka seja državnega zbora različne prekinitve. Takšne prekinitve smo izrezali iz posnetkov, saj so nepotrebne za učenje akustičnih modelov. V nadaljevanju smo analizirali razlike med rednimi in izrednimi sejami državnega zbora. Potrdila se je napoved, da redne seje pokrivajo širši spekter tematik, ter so običajno napovedane za daljše obdobje vnaprej. Nasprotno je vsebina izrednih sej osredotočena na eno samo tematiko, običajno pa so sklicane v krajšem časovnem roku. Slovar izrednih sej je praviloma manjši in bolj homogen, kot pri rednih sejah. To dejstvo olajša razvoj razpoznavalnika govora, saj je tako lažje adaptirati slovar razpoznavalnika, ter s tem zmanjšati delež besed izven slovarja. Le-ta predstavlja za pregibne jezike eno izmed glavnih ovir za dosego dobrega rezultata. Redne seje državnega zbora, vključene v učni set baze SloParl so v povprečju za približno 12% daljše kot izredne seje.

Bistveni del govorne baze za učenje akustičnih modelov so transkripcije izgovorjenega. Osnovno vodilo, ki smo mu sledili pri izdelavi transkripcij za bazo SloParl, je bilo čimbolj zmanjšati količino potrebnega ročnega dela. Tako smo kot osnovo vzeli magnetograme sej, ki so dostopni na domači strani parlamenta. Le-ti v veliki meri vsebujejo prepis govora iz razprav. V magnetogramih manjkajo oznake za efekte spontanega govora (npr.: mašila, ponovni štarti, zatikanja,...), katerih modeliranje lahko izboljša kvaliteto akustičnih modelov. Po drugi strani pa magnetogrami vsebujejo različne dodatne meta informacije (npr.: ime govorca, rezultati glasovanje, časovne oznake,...), ki niso neposredno povezane z izgovorjenim besedilom. Ime govorca, datum, številko in tip seje smo obdržali v glavi transkripcije kot meta informacijo, medtem ko smo ostanek takšnih informacij izločili iz transkripcij.

Učni nabor govornega korpusa SloParl smo razdelili v dva enako velika dela. Vsak je velik 46 ur. V prvem delu smo transkripcije pustili v takšni neobdelani obliki. Drugi polovici transkripcij smo ročno dodali časovne meje za vsako menjavo govorcev. Dodatna časovna informacija lahko izboljša kvaliteto akustičnih modelov (Lamel et al., 2002), in jo je možno relativno hitro in preprosto dodati v transkripcije.

Če želimo uporabljati razvojni in testni nabor govornega korpusa za razvoj sistema za razpoznavanje tekočega govora, potrebujemo popolne transkripcije izgovorjenega. Tako je potrebno oba nabor ročno transkribirati. Pri tem smo uporabili tri fazni pristop, ki smo ga uporabili že pri izdelavi slovenske baze BNSI Broadcast News (Žgank et al., 2004). Magnetogrami sej so služili za tvorjenje inicialne verzije transkripcij. Za obdelavo transkripcij smo uporabili pravila zapisovanja in delovno okolje (program Transcriber (Barras et al., 2001)), ki smo jih uporabljali že v projektu BNSI Broadcast News.

Da bi pokazali obseg in kompleksnost govornega korpusa slovenske baze SloParl smo opravili analizo transkripcij. Rezultati statistike so podani v tabeli 2.

Dvajset sej vključenih v slovensko bazo SloParl vsebuje 3665 menjav govorcev. V celotnem govornem korpusu so izgovorjave 255 različnih govorcev. Transkripcije 100 ur govornega materiala vsebujejo skupaj 655k besed, kjer je 37k besed različnih. Za primerjavo poglejmo obseg

Parameter	Vrednost
Menjave govorca	3665
Število govorcev	255
Število besed	655k
Število različnih besed	37k

Tabela 2: Analiza transkripcij govornega korpusa slovenske baze SloParl.

slovenske govorne baze BNSI Broadcast News (Žgank et al., 2004): ta v 36 urah govornega materiala pokriva 1565 govorcev, ki so izgovorili 268k besed (37k različnih). Približno isto število različnih besed v bazah SloParl in BNSI je verjetno posledica dejstva, da je tematika parlamentarnih razprav praviloma ožja, kot je tematika dnevnoinformativnih oddaj. Ker so politiki pogosto gostje v dnevnoinformativnih oddajah, smo primerjali oba nabora govorcev. V obeh se pojavlja 89 govorcev, kar predstavlja 34,9% govorcev v bazi SloParl. Analizirali smo tudi prekrivanje med slovarjem govornega korpusa SloParl in BNSI Broadcast News. Prekrivanje je bilo 46,3%. Glede na relativno nizek nivo prekrivanja slovarjev je možno sklepati, da bo bazo SloParl možno učinkovito uporabiti kot dopolnilo k bazi BNSI Broadcast News.

4. Tekstovni korpus

Drugi – tekstovni – del baze SloParl je namenjen učenju jezikovnih modelov razpoznavalnika tekočega govora. Kvaliteta jezikovnih modelov, ki jih uporabimo v razpoznavalniku tekočega govora za pregibne jezike je še posebej pomembna, saj ocena jezikovnega modela bistveno vpliva na doseženi rezultat. Neobdelan tekst s parlamentarnimi razpravami smo zajeli na domači strani Državnega zbora Republike Slovenije. V tekstovni korpus za izdelavo jezikovnih modelov smo vključili vse redne in izredne seje iz obdobja 1996 - 2005.

Tekst razprav dostopen na domači strani državnega zbora uporablja kodno tabelo UTF-8, ki smo jo zaradi združljivosti z ostalimi slovenskimi jezikovnimi viri spremenili v kodno tabelo ISO 8859-2. Kot je že bilo omenjeno, vsebujejo magnetogrami nekatere dodatne informacije, kot je na primer: ime govorca, rezultat glasovanja, časovne meje... Analiza, katere izmed teh informacij je smiselno ohraniti v tekstovnem korpusu za učenje jezikovnih modelov, je pokazala na sledeče parametre: ime govorca, datum, številka seje in tip seje. Te parametre smo ohranili v tekstovnem korpusu baze SloParl kot dodatno informacijo. Pričakujemo lahko, da bomo te informacije koristno uporabili med postopkom razvoja razpoznavalnika govora. Eden izmed možnih načinov uporabe teh parametrov je združevanje različnih tipov sej za zožanje tematike jezikovnih modelov. Vse ostale dodatne informacije, ki so se nahajale v magnetogramih in niso predstavljale izgovorjenega, smo izločili iz tekstovnega korpusa.

V nadaljevanju smo opravili analizo tekstovnega korpusa baze SloParl – rezultati analize so predstavljeni v tabeli 3.

V tekstovni korpus slovenske baze SloParl smo vključili 10 let parlamentarnih razprav. V tem obdobju se je odvi-

Parameter	Vrednost
Letniki	10
Število sej	188
Redne seje	69
Izredne seje	119
Število razprav	781
Število besed	23M
Število različnih besed	182k

Tabela 3: Statistika tekstovnega korpusa slovenske baze SloParl.

jalo 188 sej, od tega jih je bilo 69 rednih, preostalih 119 pa izrednih. Posamezna seja traja običajno več kot en dan, zato je skupno število razprav bistveno višje. V tem obdobju je tako bilo 781 razprav, ki smo jih vključili v tekstovni korpus. Le-ta vsebuje skupaj 23M besed, kjer je 182k besed različnih. Tiste seje državnega zbora, ki smo jih vključili v razvojni in testni nabor govornega korpusa, smo v celoti izločili iz tekstovnega korpusa.

Predstavitev novega slovenskega jezikovnega vira, baze SloParl, bomo zaključili s primerjavo z drugimi slovenskimi tekstovnimi korpusi. Tukaj je kot posebej pomembno dejstvo potrebno izpostaviti, da tekstovni korpus baze SloParl vsebuje znaten delež zapisov govorjenega jezika, ki se po svojih tipičnih lastnostih bistveno loči od pisanega besedila v časopisnih korpusih, ki so običajno v uporabi. To je še posebej pomembno pri razpoznavanju spontanega tekočega govora, kjer želimo doseči čim večjo skladnost med akustičnimi in jezikovnimi modeli. V bazi SloParl tako najdemo spontan govora v primeru poslanskih replik. Po drugi strani je bran govor prisoten v primerih predstavitve predlogov zakonov. Primerjava s časopisnim korpusom Večer pokaže, da le-ta vsebuje 105M besed, od tega 660k različnih. Če sedaj primerjamo razmerje med skupnim številom besed in številom različnih besed za oba tekstovna korpusa, lahko vidimo, da je le-to podobno, kar kaže na primerljivo kompleksnost obeh tekstovnih korpusov.

5. Zaključek

V članku smo predstavili nov slovenski jezikovni vir, bazo SloParl, ki vsebuje parlamentarne razprave Državnega zbora Republike Slovenije. Sestavljata jo govorni in tekstovni korpus. Osnovno vodilo pri izdelavi baze je bilo na stroškovno učinkovit način povečati število slovenskih jezikovnih virov, ki so namenjeni avtomatskemu razpoznavanju tekočega govora z velikim slovarjem besed.

Govorni korpus SloParl bomo uporabljali za nenadzorovano oziroma rahlo nadzorovano učenje akustičnih modelov. Tekstovni korpus je v kombinaciji s časopisnimi korpusi namenjen učenju jezikovnih modelov. Pri nadaljnjem delu se bomo osredotočili na razvoj razpoznavalnika govora za področje parlamentarnih razprav.

Zahvala

Avtorji članka se zahvaljujejo osebju Državnega zbora Republike Slovenije, ki je sodelovalo pri izvedbi projekta SloParl.

6. Literatura

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M., "Transcriber: Development and use of a tool for assisting speech corpora production", Speech Communication, Vol. 33, Issues 1-2, 5-22, 2001.

Biatov, K., Köhler, J., "Methods and Tools for Speech Data Acquisition exploiting a Database of German Parliamentary Speeches and Transcripts from the Internet", Proc. LREC 2002, Las Palmas, Španija, junij 2002.

Byrne, W., Hajic, J., Ircing, P., Khudanpur, F., McDonough, J., Peterek, N., and Psutka, J., "Large vocabulary speech recognition for read and broadcast Czech", Proc. Workshop on Text Speech and Dialog, Plzen, Češka, 1999, Lecture Notes in Artificial Intelligence, Vol. 1692.

Dreo, D., "Slovene speech data base SNABI", Dialog Man-Machine : second International Workshop, Maribor, Slovenija, 1995.

Gollan, C., Biasni, M., Kanthak, S., Schlüter R., Ney, H., "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus", Proc. ICASSP 2005, Philadelphia.

Kemp, T., Waibel, A., "Unsupervised Training Of A Speech Recognizer: Recent Experiments", Proc. Eurospeech 1999, Budimpešta, Madžarska.

Lamel, L., Gauvain, J., and Adda, G., "Lightly supervised and unsupervised acoustic model training", Computer Speech & Language, Volume 16, Issue 1, , januar 2002, 115–129.

Nouza, J., Nejedlova, D., Zdansky, J., Kolorenc, J., "Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs", Proc. ICSLP 2004, Jeju Island, Koreja.

Wessel, F., Ney, H., "Unsupervised training of acoustic models for large vocabulary continuous speech recognition". In Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italija, december 2001.

Žgank, A., Kačič, Z., and Horvat, B., "Large vocabulary continuous speech recognizer for Slovenian language", Proc. Text, speech and dialogue : 4th international conference, TSD 2001, Železna Ruda, Češka, Lecture notes in Artificial Intelligence, Vol. 2166, 242–248, Springer 2001.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B., "Acquisition and annotation of Slovenian Broadcast News database", Fourth international conference on language resources and evaluation, Lizbona, Portugalska. LREC 2004, Vol. 6, 2103–2106, 2004.

Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, Z., "SINOD - Slovenian non-native speech database", Proc. LREC 2006, Genova, Italija.

Žibert, J., Mihelič, F., (2004) Development of Slovenian broadcast news speech database. *Fourth International Conference on Language Resources and Evaluation*, Lizbona, Portugalska.