Širjenje slovarja in dvoprehodni algoritem v razpoznavalniku tekočega govora UMB Broadcast News

Gregor Donaj, Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor
gregor.donaj@uni-mb.si, kacic@uni-mb.si

Povzetek

V članku bomo predstavili nekatere najnovejše poskuse na razpoznavalniku slovenskega govora z velikim slovarjem UMB Broadcast News. Različico sistema, ki je bila predstavljena leta 2010, smo nadgradili z jezikovnimi modeli z večjimi slovarji in dvoprehodnim algoritmom, ki uporablja različne jezikovne modele v prvem in drugem prehodu. V članku primerjamo delež napačno razpoznanih besed in faktor realnega časa razpoznavanja pri različnih velikostih slovarja in različnih redih jezikovnih modelov. S predstavljeno različico sistema smo z enoprehodnim algoritmom dosegli najmanjši delež napačno razpoznanih besed 25,65%. Pokazali smo tudi, da lahko z uporabo dvoprehodnega algoritma dosežemo primerljivo uspešnost razpoznavanja v bistveno krajšem času. Prav tako nam predstavljeni rezultati služijo tudi kot smernice za nadaljevanje dela na tem področju.

Vocabulary enlargement and a two-pass algorithm for the UMB Broadcast News continuous speech recognizer

In this paper we present some recent experiments on the UMB Broadcast News large vocabulary continuous speech recognition system. We took the 2010 version and added new language models with larger vocabularies and a two-pass recognition algorithm that uses different language model in its two passes. We compare word error rates and real time factors on different vocabulary sizes and model orders. We achieved a minimum word error rate of 25.65% on a one-pass algorithm. We also show that comparable results can be achieved with significantly less time effort using a two-pass algorithm. The presented results also serve as guidelines for further work in this area.

1. Uvod

Razpoznavanje tekočega govora z velikim slovarjem je kljub napredkom tehnologije še vedno ena najzahtevnejših nalog na področju procesiranja govora. To velja tako za uspešnost kot tudi za hitrost razpoznavanja. V tem članku[1] bomo predstavili nekatere najnovejše pristope za izboljšanje uspešnosti v sistemu za razpoznavanje tekočega govora UMB Broadcast News, ki je bil razvit na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru.

Prva različica sistema UMB BN je bila predstavljena leta 2006 (Žgank & Sepesy Maučec, 2006). Zadnja predstavljena različica je iz leta 2010 (Žgank & Sepesy Maučec, 2010), v kateri so bili izboljšani akustični in jezikovni modeli. Dosežena je bila pravilnost razpoznanih besed 71,3% oz. delež napačno razpoznanih besed (NRB) 28,7%. Iz te različice smo prevzeli akustične modele, izdelali pa smo nove jezikovne modele. Namen našega dela je bil izboljšati uspešnost razpoznavanja ob čim manjšem povečanju časovne zahtevnosti.

1.1. Faktor realnega časa

Časovno zahtevnost razpoznavanja merimo s faktorjem realnega časa (RTF). To je razmerje med procesorskim časom, ki ga iskalni algoritem potrebuje, da iz nekega zvočnega segmenta razpozna izgovorjene besede, in dolžino tega segmenta. RTF je odvisen tako od uporabljene strojne in programske opreme, kot tudi od uporabljenih modelov. Nas je zanimal vpliv uporabe različnih jezikovnih modelov na RTF. Da smo lahko primerjali rezultate za naše modele, smo vse poskuse izvajali v enakih pogojih na istem strežniku.

1.2. Besede OOV

Pri povečevanju slovarja smo se osredotočili na zmanjšanje napak, ki nastanejo zaradi besed izven slovarja (OOV). Pri uporabi dvoprehodnega algoritma pa na možnost uporabe jezikovnih modelov ločeno od akustičnega razpoznavanja, ki je časovno zelo zahtevno. Slovenščina je pregibni jezik, zaradi česar vsebuje veliko več besednih oblik, kot pa drugi jeziki, kot je na primer angleščina (Rotovnik et al., 2003). Posledično se ob enaki velikosti slovarja v slovenskem besedilu pojavi več besed OOV. Kadar se v besedilu pojavi beseda, ki se ne nahaja v slovarju razpoznavalnika, je ta tudi ne more pravilno prepoznati. Posledično se pojavi napaka. Zmanjšanje deleža OOV dosegamo z večanjem slovarja. Ocenjujemo, da pri slovenščini in drugih slovanskih jezikih, kot sta češki in ruski, potrebujemo približno sedem do desetkrat večje slovarje za enako pokritost korpusa (Rotovnik et al., 2003; Zablotskiy et al., 2010; Nouza et al., 2010).

1.3. Iskalni algoritem

Iskalni algoritem (Aubert, 2002) je ključnega pomena za uspešnost razpoznavalnika. Njegova naloga je s pomočjo slovarja in akustičnih ter jezikovnih modelov najti besedne hipoteze, ki najbolje ustrezajo nekemu zvočnemu posnetku. Poznamo različne vrste iskalnih algoritmov. V tem delu smo uporabljali dvoprehodni algoritem.

V prvem prehodu uporabljamo tako akustične kot tudi bigramske in trigramske jezikovne modele. V drugem pre-

[1]Delo je bilo delno sofinancirano s štipendijo ARRS mladega raziskovalca po pogodbi 1000-10-310131

hodu pa uporabljamo le štirigramske jezikovne modele. V drugem prehodu smo delali le s tekstovnimi datotekami. Ker je obdelovanje teksta časovno veliko manj zahtevno, kot pa razpoznavanje iz zvočnega posnetka, smo upali s tem pristopom v razmeroma kratkem času doseči izboljšanje rezultatov, ki jih dobimo po prvem prehodu.

1.4. Struktura članka

V nadaljevanju bomo najprej predstavili uporabljene govorne in tekstovne vire za izdelavo modelov. V tretjem poglavju članka bomo predstavili način gradnje slovarjev in jezikovnih modelov. V četrtem poglavju bo predstavljeno delovanje uporabljenega dvoprehodnega razpoznavalnega algoritma. Rezultati razpoznavanja bodo predstavljeni v petem poglavju. Najpomembnejši zaključki in nekatere smernice za možnost nadaljevanja dela bomo predstavili v šestem poglavju.

2. Uporabljeni viri

Za gradnjo modelov sta bila uporabljena slovenska govorna baza BNSI Broadcast News (Žgank et al., 2008/2) in slovenski referenčni jezikovni korpus FidaPLUS (Arhar & Gorjanc, 2007).

Baza BNSI je sestavljena iz 36 ur govornega materiala, ki je bil zbran iz različnih informativnih televizijskih oddaj RTV Slovenije. Največji del baze so učni podatki, na katerih so naučeni akustični modeli. Za učenje modelov smo uporabili orodje HTK (HTK, 2010). Razvojna in testna množica obsegata po slabe 3 ure materiala. Vse poskuse razpoznavanja smo izvajali na testni množici. Razvojna množica je namenjena optimiziranju parametrov modela, ki smo jih prevzeli iz prejšnje verzije. Baza vsebuje tudi tekstovni del. Ta obsega približno 11 milijonov besed, vendar pa ga pri izdelavi modelov nismo uporabljali.

Za gradnjo slovarjev in jezikovnih modelov smo uporabljali korpus FidaPLUS, ki vsebuje različna slovenska besedila s skupno 621 milijoni besed. Korpus je lematiziran in vsebuje morfosintaktične oznake, ki pa jih nismo uporabljali.

3. Širitev slovarja in gradnja jezikovnih modelov

Korpus FidaPLUS smo najprej obdelali tako, da smo iz njega izluščili le besede. Števila, okrajšave in druge besede, ki so vsebovale številke ali pa posebne znake, smo nadomestili s posebnimi oznakami. Odpravili smo vsa naglasna znamenja in podobne oznake ob črkah.

Slovarje smo zgradili tako, da smo v njih dodajali besede korpusa v vrstnem redu glede na njihovo pogostost. Kadar smo dosegli želeno velikost slovarja smo dodali še vse besede, ki se pojavijo enako pogosto, kot zadnja dodana beseda. Pri tem smo izbrali različne želene velikosti slovarjev od 60.000 besed do 300.000 besed. V slovarjih sta zraven besed le še oznaki za začetek in konec stavka. Vse ostale oznake smo izključili iz slovarja. Končni slovarji so zaradi načina njihove gradnje bili nekoliko večji od želenih velikosti.

Modele bomo kasneje vrednotili na testni množici BNSI. Za vsak slovar smo izračunali delež besed v testni množici, ki se ne pojavijo v slovarju. Tabela 1 prikazuje velikosti slovarjev in deleže OOV. Ti so primerljivi z rezultati na češkem jeziku (Nouza et al., 2010), kjer je bilo ugotovljeno tudi izboljšanje uspešnost razpoznavanja primerljivo z zmanjšanjem deleža besed OOV. Podobna izboljšanja pričakujemo tudi za slovenščino.

Tabela 1: Velikosti slovarjev in deleži besed OOV na testni množici.

Slovar	Velikost	OOV [%]
60k	60.022	6,94
100k	100.189	3,44
150k	150.285	2,24
200k	201.034	1,64
250k	251.352	1,29
300k	301.357	1,02

Za vse velikosti slovarjev smo zgradili bigramske (2g), trigramske (3g) in štirigramske (4g) modele. Uporabljali smo Good-Turingovo glajenje in sestopanje po Katz-u. Posebne oznake, ki smo jih uporabljali pri obdelavi korpusa nismo vključili v modele, saj se uporabljajo le pri pisanih besedilih in zato pri modeliranju govorjenega besedila niso uporabne. Med modeliranjem se obnašajo kot neznane besede.

4. Dvoprehodni algoritem

V izločanju značilk iz zvočnih posnetkov smo uporabljali mel kepstralne koeficiente (MFCC)(Biem et al., 2005) in energijo signala ter prve in druge odvode. Uporabljali smo 26 mel filtrov. Značilke smo izločali v oknih dolžine 32 ms in v presledkih 10 ms. Uporabljeni akustični modeli so medbesedni trigrafemski zvezni HMM s 16 Gaussovimi porazdelitvami in vezanimi stanji.

4.1. Prvi prehod

Prvič smo na razpoznavalniku UMB Broadcast News uporabili dvoprehodni algoritem. V prvem prehodu smo uporabljali orodje HDecode (HTK, 2010). Med razpoznavanjem se uporablja sinhroni iskalni algoritem z Viterbijevo aproksimacijo in snopovnim omejevanjem. Prvi prehod smo izvajali z vsemi predstavljenimi velikostmi slovarja ter bigramskimi in trigramskimi modeli.

4.2. N-najboljših seznami

V prvem prehodu algoritma smo dobili za vsak segment govora dva rezultata. Prvi je bil najboljša hipoteza, ki se lahko neposredno ovrednoti (izračunamo število napačno razpoznanih besed). Iz najboljših hipotez smo kasneje izračunali uspešnost razpoznavanja prvega prehoda.

Istočasno pa nam je iskalni algoritem vrnil tudi besedni graf (besedno mrežo), v katerem je predstavljen iskalnih prostor algoritma ob zaključku segmenta. Iz nje lahko razberemo več različnih besednih hipotez, ki jih je iskalni algoritem vrednotil med delovanjem. Za vsako besedo v grafu imamo podan čas začetka in konca ter verjetnosti akustičnega in jezikovnega modela. Iz besednih grafov lahko razberemo več različnih hipotez za celotne segmente.

Uporabili smo orodje s katerim smo iz vseh besednih mrež izpisali sezname 1000 najboljših hipotez. Te sezname smo kasneje uporabljali v drugem prehodu algoritma. V nekaterih primerih je bil iskalni prostor ob delovanju algoritma tako majhnem, da ni bilo možno tvoriti 1000 hipotez. V teh primerih smo dobili manjše sezname. Za vsako hipotezo imamo v seznamih podane vse podatke s katerimi se hipoteze ocenjujemo: število besed ter verjetnosti akustičnega in jezikovnega modela.

4.3. Drugi prehod

Osnovna ideja predstavljenega dvoprehodnega algoritma je ta, da s kompleksnejšimi modeli ponovno ovrednotimo vse hipoteze. Uporabljali smo orodje SRI Language Modeling Toolkit (Stolcke, 2002).

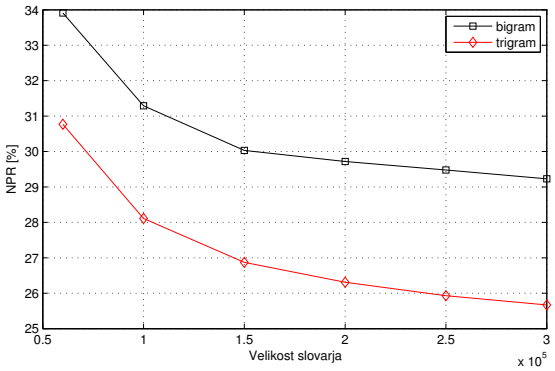
V drugem prehodu smo uporabljali le jezikovne modele. Verjetnost akustičnega modela smo prevzeli iz prvega prehoda. Nove verjetnosti jezikovnih modelov pa smo dobili z uporabo štirigramskih modelov. Ko smo ocenjevali hipoteze iz prvega prehoda smo vedno uporabljali model, ki je bil zgrajen na istem slovarju kot model iz prvega prehoda. Potem, ko se vse hipoteze ponovno ocenijo, se izbere nova najboljša hipoteza, ki se uporablja v vrednotenju uspešnosti razpoznavanja.

5. Rezultati eksperimentov

Uspešnost izdelanih modelov in algoritma smo vrednotili na ročno segmentirani testni množici baze BNSI. Vse rezultate podajamo v deležu napačno razpoznanih besed. To je razmerje med vsoto zamenjanih, vrinjenih in izbrisanih besed ter številom vseh besed v testni množici.

Najprej smo ovrednotili uspešnost prvega prehoda algoritma za različne velikosti slovarja. Rezultati so podani na sliki 1. Iz grafa je razvidno manjšanje števila napak pri večanju slovarja tako pri uporabi bigramskega kot trigramskega modela. Pri primerjavi rezultatov med slovarjema 60k in 300k ugotovimo, da je bilo z bigramskim modelom doseženo zmanjšanje deleža napak za 4,67%, pri trigramskem modelu pa 5,09%. Ta izboljšanja so primerljiva z zmanjšanjem deleža besed izven slovarja, ki znaša 5,92%. Ta primerljivost je bila pričakovana, saj smo predvidevali, da bomo z večanjem slovarja izločili napake, ki nastanejo zaradi besed izven slovarja. Majhen delež preostalih besed izven slovarja in potek grafov na sliki pa kažeta na možnost, da z dodatnim večanjem slovarja nad 300k ne bomo več dosegli bistvenih izboljšav.

V nadaljevanju smo se osredotočili na modele s slovarjema 60k in 300k. V tabeli 2 so predstavljeni rezultati napačno razpoznanih besed in faktorja realnega časa. Iz podatkov vidimo, da povečanje slovarja iz 60k na 300k prinese 4,67% oz. 5,09% zmanjšanje deleža napačno razpoznanih besed. Izboljšanje pri zamenjavi bigramskega modela s trigramskim pa prinese 3,15% oz. 3,57% zmanjšanje deleža napačno razpoznanih besed. Iz podatkov o faktorju realnega časa lahko vidimo, da se algoritem ob uporabi večjega slovarja upočasni približno za faktor 2, ob zamenjavi bigramskega modela s trigramskim pa za faktor 3. Iz teh ugotovitev lahko povzamemo, da ima uporaba večjega slovarja tako večji doprinos k uspešnosti razpoznavanja kot



Slika 1: Uspešnost prvega prehoda.

tudi manjše povečanje časovne zahtevnosti kot pa zamenjava bigramskih modelov s trigramskimi.

Dobljeni rezultati za deleže besed OOV in rezultate razpoznavanja so primerljivi z sistemom, opisanem v (Nouza et al., 2010), za razpoznavanje češkega jezika, ki je po svojih lastnostih podoben slovenskemu.

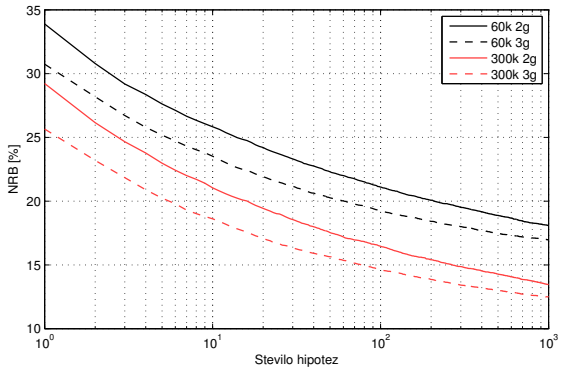
V naših poskusih smo uporabljali le ročno segmentirano testno množico. Izkušnje iz preteklih različic sistem pa kažejo, da z uporabo avtomatske segmentacije izgubimo približno 2 do 3 odstotke uspešnosti razpoznavanja (Žgank et al., 2008/1).

Tabela 2: Rezultati prvega prehoda.

Model	NRB [%]	RTF
60k 2g	33,89	6,29
60k 3g	30,74	18,46
300k 2g	29,22	12,66
300k 3g	25,65	37,09

Preden smo vrednotili razpoznavanje v drugem prehodu, smo si podrobneje pogledali hipoteze v N-najboljših seznamih. Na sliki 2 so prikazani deleži napak v N-najboljših seznamih, ki jih dobimo tako, da vzamemo najboljših n hipotez, pogledamo število napak v vsaki hipotezi in kot rezultat podamo najmanjše število napak. Tako dobimo pričakovano izboljšanje rezultata, če bi v naslednjem prehodu algoritma imeli na voljo idealni razpoznavalnik, ki bi vedno znal izbrati najboljšo hipotezo. Iz slike 2 je razvidno padanje števila napak z večanjem števila hipotez. Pri 1000 hipotezah se število napak razpolovi. Iz grafa je tudi razvidno, da je padanje števila napak hitrejše pri rezultatih, ki so bili dobljeni z uporabo bigramskega modela.

Zadnji rezultati, ki smo jih vrednotili se nanašajo na dvoprehodni algoritem. Podani so v tabeli 3. Rezultate prvega prehoda iskalnega algoritma pri modelih iz tabele 2 (prvi stolpec) smo ponovno vrednotili s štirigramskimi modeli. Izračunali smo nov delež napačno razpoznanih besed (tretji stolpec) in razliko glede na rezultat prvega prehoda (četrti stolpec). Izboljšanje rezultatov se kaže v primerih, kjer smo v prvem prehodu uporabljali bigramski jezikovni model. Pri teh primerih smo poskuse ponovili tudi s trigramskimi jezikovnimi modeli v drugem prehodu, ven-



Slika 2: Napake v N-najboljših seznamih.

dar ni opaznih razlik do podanih rezultatov uspešnosti in faktorja realnega časa pri uporabi štirigramskega modela. Iz rezultatov lahko predvidevamo, da dodatno višanje reda običajnih jezikovih modelov ne bo prineslo dodatnih izboljšav.

Tabela 3: Rezultati drugega prehoda.

1. prehod	2. prehod	NRB [%]	ΔNRB [%]	RTF
60k 2g	60k 4g	31,01	-2,88	0,02
60k 3g	60k 4g	30,73	-0,01	0,02
300k 2g	300k 4g	25,85	-3,37	0,03
300k 3g	300k 4g	25,64	-0,01	0,03

V zadnjem stolpcu so podani še faktorji realnega časa drugega prehoda. Ker v drugem prehodu delamo le z omejenim številom hipotez in uporabljamo le jezikovne modele, so ti faktorji zelo majhni v primerjavi s faktorji prvega prehoda. Skupni faktor realnega časa za dvoprehodni algoritem dobimo tako, da faktorja obeh prehodov seštejemo. Čas, ki je potreben za pretvorbo besednih grafov v sezname N-najboljših hipotez je v primerjavi s časom delovanja algoritma zanemarljiv. Za hitrost delovanja algoritma je torej pomemben predvsem prvi prehod. Vidimo, da smo z uporabo bigramskega modela v prvem prehodu in štirigramskega v drugem prehodu dosegli uspešnost razpoznavanja primerljivo (razlika 0,12% oz. 0,2%) z uspešnostjo enoprehodnega algoritma s trigramskim modelom, ki pa zahteva trikrat več časa za razpoznavanje.

6. Zaključek

V članku smo predstavili nekatere rezultata iz trenutnega razvoja razpoznavalnika UMB Broadcast News. Z uporabo razširjenega slovarja smo uspeli zmanjšati napako razpoznavanja na 25,65%. Uporaba dvoprehodnega algoritma pa nam je omogočila doseganje primerljivih rezultatov s trikrat krajšim časom razpoznavanja.

Predstavljeni rezultati nakazujejo, da z dodatnim večanjem slovarja in vpeljavo običajnih jezikovnih modelov višjih redov, kot pa smo jih sedaj uporabljali, ni več mogoče pričakovati bistvenih izboljšanj v uspešnosti razpoznavalnika.

Izboljšanja lahko pričakujemo s kombiniranjem predstavljenih pristopov in nekaterih drugih že pripravljenih izboljšav razpoznavalnika, kot so interpolirani jezikovni in izboljšani akustični modeli. Predstavljen sistem dvoprehodnega algoritma pa nam omogoča začetek raziskovanja na zapletenejših jezikovnih modelih, ki bi jih zgradili z uporabo morfosintaktičnih oznak iz korpusa FidaPLUS.

7. Literatura

Arhar Š., Gorjanc, V. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2):95–110.

Aubert., X.L. 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16(1):89–114.

Biem A., Mcdermott A., Katagiri, S. 2005. A discriminative filter bank model for speech recognition. In *ESCA Eurospeech 2005*, Lizbona, Portugalska.

HTK domača stran, http://htk.eng.cam.ac.uk

Nouza, J., Zdansky, J., Cerva, P., Silovsky, J. 2010. *Development of Multimodal Interfaces: Active Listening and Synchrony*, pogl. Challenges in speech processing of slavic languages (case studies in speech recognition of czech and slovak) Springer Berlin / Heidelberg

Rotovnik, T., Sepesy Maučec, M., Kačič Z. 2003. Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Communication*, 49(6):437–452.

Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. In *International Conference on Speech and Language Processing*, Denver, Colorado.

Zablotskiy, S., Zablotskaya K., Minker W. 2010. Some approaches for russian speech recognition. In *Sixth Internation Conference on Intelligent Environments*, Kuala Lumpur, Malezija.

Žgank, A., Sepesy Maučec, M. 2006. Osnovna zgradba razpoznavalnika slovenskega tekočega govora UMB Broadcast News. *Jezikovne tehnologije 2006*, Ljubljana, Slovenija

Žgank, A., Kos, M., Kotnik, B., Sepesy Maučec, M., Rotovnik, T., Kačič, Z. 2008. Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast News *Jezikovne tehnologije 2008*, Ljubljana, Slovenija

Žgank, A., Verdonik, D., Kačič, Z. 2008. Slovenska baza BNSI Broadcas News za razpoznavanje tekočega govora. *Elektrotehniški vestnik*, 75(3):85–90.

Žgank, A., Sepesy Maučec, M. 2010. Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Jezikovne tehnologije 2010*, Ljubljana, Slovenija