Vpliv izbire atributov na točnost naglaševanja nepoznanih slovenskih besed

Tea Robič, Tomaž Šef

Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova 39, SI-1000 Ljubljana
tea.robic@ijs.si, tomaz.sef@ijs.si

Povzetek

Naglaševanje besed je sestavni del vsakega sistema za sintezo govora. Za slovenski jezik je ta naloga nekoliko težja kot pri večini ostalih jezikov, saj imajo slovenske besede prosto mesto naglasa – to pomeni, da je naglas lahko na katerem koli zlogu besede. Zato smo pri naglaševanju nepoznanih besed uporabili strojno učenje, s katerim določimo mesto naglasa s pomočjo atributov, ki opisujejo različne lastnosti besede in opazovanega grafema. V prispevku primerjamo klasifikacijsko točnost pri naglaševanju slovenskih besed glede na različne nabore atributov pri strojnem učenju. Poskuse smo opravili z atributi, ki nosijo informacijo o osnovnih lastnostih besede (kot so število zlogov v besedi, predpona, pripona, kontekst opazovanega grafema in druge), atributi, ki poleg osnovnih lastnosti upoštevajo še oblikoslovne oznake besede in atributi, ki poleg vseh omenjenih lastnosti upoštevajo še mesto zloga šteto od konca besede. Rezultati kažejo, da upoštevanje oblikoslovnih oznak in dodatnega atributa pomaga k večji klasifikacijski točnosti.

The Influence of Attribute Choice on the Accentuation Accuracy of Unknown Slovene Words

Accentuation of words is a basic task in the construction of any speech synthesis system. This task is particularly difficult in languages where lexical stress can be located almost arbitrarily on every syllable in the word, such as in the Slovene language. Therefore, we use machine learning methods to create accentuation rules. In this paper we compare the classification accuracy of accentuation when using different sets of attributes. We experiment with attributes, that define the basic properties of the word (such as the number of syllables in the word, prefix, suffix, the context of the observed grapheme, etc.), and attributes that consider (beside all mentioned attributes) also the position of the observed syllable counted from the end of the word. As shown by the results, the best classification accuracy is achieved when all attributes are considered. This suggests that the morphological information and the application of the additional attribute help in the task of accentuation of Slovene words.

1. Uvod

Pretvorba grafemov v foneme je pomembna naloga v vsakem sistemu, ki želi omogočati samodejno sintezo govora. Lahko jo opišemo kot preslikavo, ki črkovnemu zapisu besede priredi njen fonemski zapis. V slovenskem jeziku lahko (v primerjavi z večino drugih jezikov) takšno pretvorbo izvedemo relativno enostavno, ko poznamo naglas besede. Vendar je ravno naglaševanje besed v slovenskem jeziku zelo zahtevna naloga, saj zanjo ni preprostih pravil.

Za slovenski jezik je značilno prosto mesto naglasa. Posamezna beseda ima lahko različno število naglašenih mest. Tako ločimo besede brez naglasa (klitike), besede z enim naglasom (večina besed) in besede z več naglasi (nekatere sestavljenke, zloženke in sklopi). Mesto naglasa je določeno za vsako besedo posebej in velja, da se ga naučimo hkrati z učenjem jezika. Poleg tega se lahko posamezna besedna oblika naglašuje na več različnih načinov – to so t.i. homografi. Na njihovo pravilno naglaševanje in izgovarjavo lahko sklepamo le iz konteksta. Takšne besedne oblike se med seboj ločijo po besedni vrsti, spolu, sklonu, številu ali pa le po pomenu.

Dosedanje metode učenja izgovarjave nepoznanih slovenskih besed temeljijo na predpostavki, da se vsa potrebna informacija v celoti nahaja v nizu znakov, ki sestavljajo besedo (Škrjanc in sod., 2002; Šef in Gams, 2004). Ker sta pri slovenščini mesto in tip dinamičnega naglasa odvisna tudi od oblikoslovnih karakteristik besede, za pravilno izgovarjavo besede potrebujemo tudi to informacijo. V tem prispevku preizkušamo model naglaševanja, pri katerem imamo za vsako besedo na voljo tudi njene oblikoslovne oznake (besedna vrsta, spol, sklon, število, oseba, čas in stopnja), ki smo jih dobili s pomočjo besedne in stavčne analize. Zanima nas, ali le-te pripomorejo k večji klasifikacijski točnosti. Poleg tega raziskujemo vpliv dodatnega atributa, ki zaznamuje mesto zloga v besedi šteto od konca besede, na klasifikacijsko točnost.

V naslednjem razdelku opisujemo podatke, ki smo jih imeli na voljo za poskuse ter način pridobivanja le-teh. Sledi prikaz atributov za strojno učenje, ki smo jih izluščili iz danih podatkov. V tretjem razdelku navajamo metodologijo, s katero smo se lotili poskusov. Uporabili smo različne metode za strojno učenje. V četrtem razdelku prikazujemo in interpretiramo dobljene rezultate. Ugotovitve zberemo v zaključku.

2. Podatki

Predpogoj za temeljito analizo naglašenosti je ustrezno velik fonetični slovar, ki vsebuje tudi oblikoslovne oznake. Tak slovar mora obsegati vse dopustne izgovarjave posamezne besede, in to v vseh njenih pojavnih oblikah.

2.1. Pridobivanje podatkov

Slovar slovenskega knjižnega jezika (SSKJ) vsebuje le besede v njihovih osnovnih oblikah, zato smo morali zgraditi nov fonetični slovar v elektronski obliki. Ta vsebuje

Tabela 1: Število primerov v učnih in testnih množicah pri trikratnem prečnem preverjanju.

	A	E	I	O	U	R	vsi
Učna množica 1	348036	326772	355327	247590	64285	19141	1361151
Testna množica 1	167688	157842	170537	122299	29321	9020	656707
Učna množica 2	340131	317148	347907	246674	62464	18387	1332711
Testna množica 2	175593	167466	177957	123215	31142	9774	685147
Učna množica 3	343281	325308	348494	245514	60463	18794	1341854
Testna množica 3	172443	159306	177370	124375	33143	9367	676004
Skupaj	515724	484614	525864	369889	93606	28161	2017858

okoli 600.000 besednih oblik, kar ustreza 20.000 lemam. Kot osnovo smo uporabili MULTEXT-East leksikon (Erjavec, 1998), ki smo ga dopolnili s podatki o mestu in tipu tako dinamičnega kot tonemskega naglasa. Poleg tega smo dodali popolne fonetične zapise besed, za katere uporabljena grafemsko-fonemska pravila ne veljajo. Večino dela smo opravlili avtomatično z uporabo v ta namen razvitega morfološkega analizatorja (ca. 50.000 vrstic programske kode v jeziku C) in elektronske verzije SSKJ-ja. Takšna določitev mesta naglasa je bila neuspešna v približno 0,2 % primerov. Poleg tega je algoritem predlagal, da dodatno preverimo še nekaj manj kot 1 % besed. V vseh teh primerih smo delo opravili ročno. Na koncu smo še enkrat pregledali celoten slovar.

Za izgradnjo domene z atributi smo uporabili 536.431 besed. Ohranili smo večkratne ponovitve posameznih besednih oblik z enako izgovarjavo, a različno oblikoslovno oznako. Kot rezultat smo dobili 2.017.858 zlogov. Te smo razdelili v tri skupine tako, da smo vse besede z isto lemo postavili v isto skupino. Nato smo eno skupino vzeli za testno množico, ostali dve pa združili v učno množico. Postopek smo ponovili trikrat (vsakič smo za testno množico vzeli drugo skupino) in dobili tri učne in tri testne množice za trikratno prečno preverjanje (tabela 1).

Vsako učno in testno množico smo nadalje razdelili na šest podmnožic – po eno za vsak samoglasnik in soglasnik 'r'. Tako smo dobili 18 učnih podproblemov.

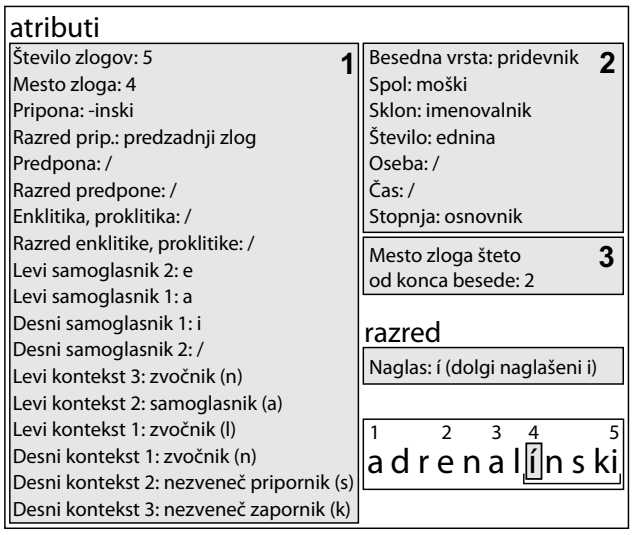
2.2. Izbira atributov

Za strojno učenje naglaševanja grafemov smo imeli na voljo 75 atributov, ki so prikazani na sliki 1:

- število zlogov v besedi (1 atribut);
- mesto opazovanega zloga v besedi (1 atribut);
- prisotnost predpone in pripone v besedi ter razred predpone in pripone (4 atributi);
- enklitika in proklitika ter razred enklitike in proklitike (2 atributa);
- samoglasniki, ki nastopajo (eno ali dve mesti) levo in desno od opazovanega samoglasnika (4 atributi);
- levi in desni kontekst opazovanega samoglasnika (trije znaki levo in desno od opazovanega samoglasnika), pri čemer vsak znak opišemo z devetimi atributi (54 atributov);
- oblikoslovne oznake besede (besedna vrsta, spol, sklon, število, čas, oseba in stopnja) (7 atributov);
- mesto opazovanega zloga v besedi šteto od konca besede (1 atribut);
- tip naglasa na izbranem samoglasniku (1 atribut) – to je razred, ki ga napovedujemo.

Želeli smo ugotoviti, s katerimi atributi dobimo najboljše rezultate. Zato smo poskuse opravili s tremi različnimi nabori atributov:

1. *Osnovni atributi.* Upoštevali smo 67 atributov, ki predstavljajo osnovne podatke o opazovanem zlogu in besedi, kot so npr. število zlogov, mesto zloga, informacija o predponi in priponi, podatki o kontekstu opazovanega grafema in drugi. Na sliki 1 so prikazani pod številko 1.
2. *Oblikoslovne oznake.* K osnovnim atributom smo dodali še 7 atributov, ki opisujejo oblikoslovne oznake opazovane besede (74 atributov pod številkama 1 in 2 na sliki 1).
3. *Dodatni atribut.* Vzeli smo vseh 75 atributov – poleg osnovnih atributov in oblikoslovnih oznak smo dodali



Slika 1: Skupine atributov za četrti samoglasnik besede *adrenalinski*.

še atribut, ki pove mesto opazovanega zloga v besedi šteto od konca besede. Čeprav je to podatek, ki ga lahko razberemo iz množice že uporabljenih atributov (mesto zloga šteto od konca besede = število zlogov v besedi − mesto zloga +1), tega izračuna metode strojnega učenja niso sposobne narediti same.

Ker smo imeli opravka s trikratnim prečnim preverjanjem na šestih podmnožicah zlogov in tremi nabori atributov, je to pomenilo kar 54 nalog učenja.

3. Poskusi

Za vsako od 54-ih nalog naglaševanja grafemov smo uporabili različne algoritme strojnega učenja, ki so implementirani v sistemu Weka 3.4.1 (Witten in Frank, 2000). Pri izboru algoritmov smo bili zaradi velikega števila atributov in primerov omejeni le na nekaj različnih pristopov. Čeprav smo uporabljali za sedanje razmere dokaj zmogljiv računalnik z 1 GB pomnilnika, so bili poskusi dolgotrajni. Tako v nadaljevanju opisujemo samo rezultate tistih metod, ki so bile dovolj učinkovite, da smo lahko z njimi opravili vseh 54 poskusov.

Uporabljene metode lahko glede na pristop reševanja razdelimo v tri skupine: metode, ki uporabljajo naivni Bayesov klasifikator, metode, ki gradijo odločitvena drevesa, in meta metode, ki klasificirajo primere glede na napovedi osnovnih klasifikatorjev.

Metode, ki uporabljajo naivni Bayesov klasifikator:

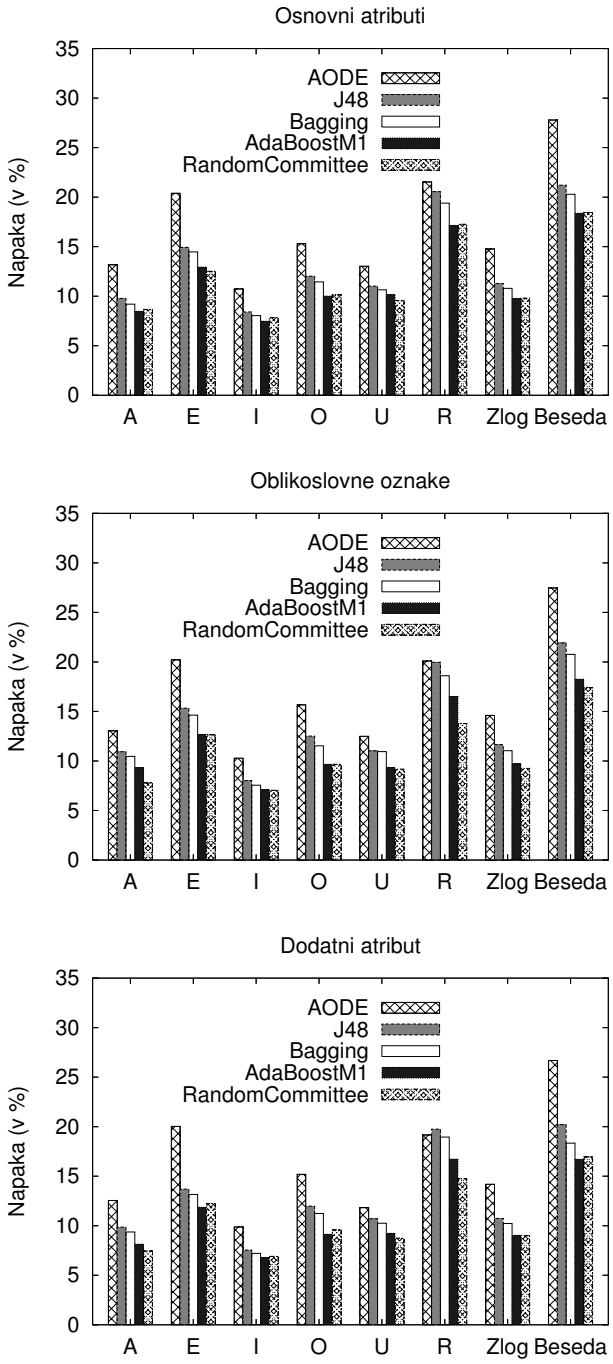
- *NaiveBayes* je metoda, ki uporablja naivni Bayesov klasifikator – predvideva, da so vsi atributi med sabo neodvisni (John in Langley, 1995).
- *AODE* uporablja več modelov, podobnih naivnemu Bayesovemu klasifikatorju, ki pa imajo šibkejše predpostavke o neodvisnosti. Ta metoda je zato navadno bolj natančna od metode NaiveBayes (Webb in sod., 2002).

Metode, ki gradijo odločitvena drevesa:

- *ADTree* generira alternirajoče odločitveno drevo (Freund in Mason, 1999).
- *RandomTree* generira odločitveno drevo, pri gradnji katerega upošteva le k naključno izbranih atributov za vsako vozlišče.
- *REPTree* zgradi odločitveno drevo glede na informacijski prispevek in drevo na koncu odreže.
- *J48* je metoda, ki zgradi odločitveno drevo C4.5 (Quinlan, 1993).

Meta metode:

- *Bagging* generira t osnovnih modelov za klasifikacijo in nato posamezne primere klasificira glede na večinsko mnenje osnovnih modelov (Breiman, 1996).
- *AdaBoostM1* implementira algoritem boosting. Deluje podobno kot metoda bagging, le da vsak od t osnovnih modelov za klasifikacijo dopolnjuje prejšnjega (Freund in Schapire, 1996).



Slika 2: Rezultati izbranih metod strojnega učenja za tri različne nabore atributov.

- *RandomCommittee* generira skupek t osnovnih klasifikatorjev, ki so zgrajeni na enakih podatkih, a inicializirani z različnim naključnim številom. Klasifikacija te meta metode je kar povprečje klasifikacij osnovnih klasifikatorjev.

Vse metode so uporabljale privzete nastavitve sistema Weka. Pri meta metodah smo parameter t nastavili na 10. Osnovni klasifikator pri metodah Bagging in AdaBoostM1 je bilo odločitveno drevo J48, pri metodi RandomCommittee pa odločitveno drevo RandomTree.

Tabela 2: Primerjava klasifikacijskih napak glede na nabor atributov.

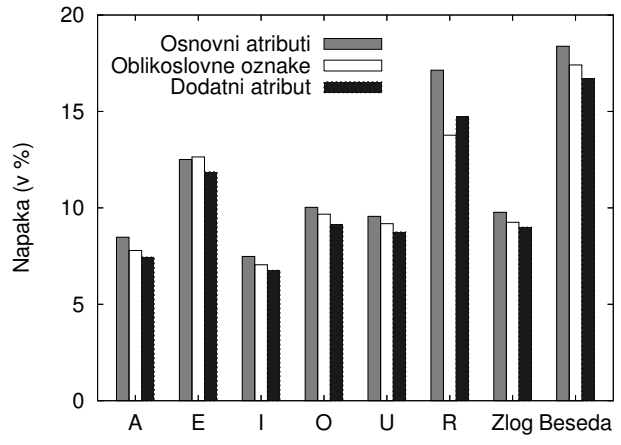
	A	E	I	O	U	R	Zlog	Beseda
Osnovni atributi	8,48 %	12,51 %	7,48 %	10,03 %	9,56 %	17,14 %	9,77 %	18,38 %
Oblikoslovne oznake	7,79 %	12,64 %	7,05 %	9,67 %	9,18 %	13,77 %	9,25 %	17,41 %
Dodani zlog	7,45 %	11,87 %	6,76 %	9,14 %	8,75 %	14,73 %	9,00 %	16,94 %

4. Rezultati

Na sliki 2 so prikazani grafi povprečnih napak metod pri naglaševanju s trikratnim prečnim preverjanjem. Poskuse smo opravili za vse samoglasnike in črko 'r' ter iz dobljenih rezultatov izračunali napako klasifikacije še za zlog in besedo.

Na grafih na sliki 2 so zaradi boljše preglednosti prikazane samo najboljše metode. Odločitvena drevesa ADTree, RandomTree in REPTree dobijo slabše rezultate kot J48, zato na sliki njihovi rezultati niso narisani. Zaradi enakega razloga smo iz slike izpustili tudi rezultate metode NaiveBayes, ki se je odrezala slabše kot metoda AODE. Najboljše rezultate so po pričakovanjih dobile meta metode. Najbolje sta se izkazali metodi RandomCommittee in AdaBoostM1.

Kot lahko vidimo iz slike 3 in tabele 2, z upoštevanjem oblikoslovnih oznak izboljšamo klasifikacijsko točnost za 0,5 % na nivoju zloga in skoraj 1 % na nivoju besede. Upoštevanje dodatnega atributa, ki pove mesto zloga v besedi šteto od konca besede, zmanjša napako klasifikacije za dodatna 0,2 % na nivoju zloga in 0,5 % na nivoju besede. Čeprav dodatni atribut poslabša točnost napovedi za soglasnik 'r' za skoraj 1 %, je število primerov za 'r' tako nizko (tabela 1), da se to na nivoju zloga oz. besede skoraj ne pozna.



Slika 3: Primerjava najboljših rezultatov za tri nabore atributov.

5. Zaključek

Naglaševanje nepoznanih besed v slovenskem jeziku predstavlja eno najzahtevnejših nalog pri sintezi govora. Za naglaševanje s čim večjo točnostjo potrebujemo poleg osnovnih informacij o grafemih in besedah tudi oblikoslovne oznake besede (besedna vrsta, spol, sklon, število, oseba, čas in stopnja). S poskusi smo pokazali, da dobimo boljše rezultate, če poleg osnovnih atributov upoštevamo še oblikoslovne oznake besede, še boljše pa, če upoštevamo tudi atribut, ki nam pove mesto zloga v besedi šteto od konca besede.

6. Literatura

L. Breiman. 1996. Bagging predictors. *Machine Learning*, 2(24):123–140.

T. Erjavec. 1998. The MULTEXT-East Slovene Lexicon. V: *Zbornik sedme Elektrotehniške in računalniške konference ERK'98*, zvezek B, str. 189–192.

Y. Freund in L. Mason. 1999. The alternating decision tree learning algorithm. V: *Proceeding of the Sixteenth International Conference on Machine Learning*, str. 124–133, Bled, Slovenia.

Y. Freund in R. E. Schapire. 1996. Experiments with a new boosting algorithm. V: *Proc International Conference on Machine Learning*, str. 148–1556. Morgan Kaufmann, San Francisco.

G. H. John in P. Langley. 1995. Estimating continuous distributions in bayesian classifiers. V: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intellingence*, str. 338–345. Morgan Kaufmann, San Mateo.

T. Šef in M. Gams. 2004. Data mining for creating accentuation rules. *Applied Artificial Intelligence*, 18(5):395–410.

M. Škrjanc, T. Šef in M. Gams. 2002. Using decision trees for accentuation in the slovenian language. V: *Proceedings STAIRS 2002, STarting Artificial Intelligence Researchers Symposium*, str. 135–144. Ios Press, Ohmsha.

R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

G. I. Webb, J. Boughton in Z. Wang. 2002. Averaged one-dependence estimators: Preliminary results. V: *Proceedings of the Australasian Data Mining Workshop (ADM 02)*, str. 65–73, Canberra. University of Technology, Sydney.

I. H. Witten in E. Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco. http://www.cs.waikato.ac.nz/˜ml/index.html.