Naglaševanje nepoznanih besed pri sintezi slovenskega govora

Tomaž Šef, Matjaž Gams, Maja Škrjanc

Inštitut Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenija
{tomaz.sef, matjaz.gams, maja.skrjanc}@ijs.si

Povzetek

V članku je predstavljen dvostopenjski model naglaševanja nepoznanih slovenskih besed, ki je uporabljen v sistemu Govorec za sintezo slovenskega govora. Najprej za vsak samoglasnik in soglasnik 'r' v besedi pogledamo, če je naglašen; za naglašene glasove določimo tudi tip naglasa. To naredimo s pomočjo odločitvenih dreves, ki smo jih dobili z metodami strojnega učenja. Sledi popravljanje tako dobljenih rezultatov glede na število naglasov v besedi in dolžino besede. Učno in testno množico smo dobili s pomočjo MULTEXT-East leksikona, ki smo ga dopolnili s podatki o mestu in tipu tako dinamičnega kot tonemskega naglasa. Dobljeni rezultati so znatno boljši od do sedaj uporabljenih ročno pridobljenih pravil. Eksperimenti so potrdili tezo, da je naglaševanje besed v slovenskem jeziku precej kompleksen problem, ki ga z relativno preprostimi pravili ni mogoče učinkovito rešiti.

1. Uvod

Mesto naglasa predstavlja zlog, na katerem ima beseda jakostno ali tonemsko izrazitost. Za razliko od nekaterih drugih jezikov (npr. francoščina - stalno mesto naglasa, na zadnjem zlogu; hrvaščina - delno omejeno mesto naglasa, zadnji zlog ni nikoli naglašen) je za slovenski jezik 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).

Za naglaševanje v slovenskem jeziku ni nikakršnih preprostih pravil. Mesto naglasa je določeno za vsako besedo posebej in velja, da se ga naučimo hkrati z učenjem jezika. Poleg tega velja omeniti, da se lahko posamezna besedna oblika naglašuje na več različnih načinov. To so tako imenovani 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.

Do sedaj so se v vseh sintetizatorjih slovenskega govora poleg relativno skromnih slovarjev izgovarjav (nekaj deset tisoč najpogostejših besed) uporabljala še zelo preprosta pravila, ki so temeljila na nekaj seznamih (breznaglasnice - enklitike in proklitike; pripone in predpone - večinoma nenaglašene; začetnice - večinoma naglašene; končaji besed z značilnimi naglasnimi mesti) (Toporišič, 1984) in statistikah, ki so podajale verjetnosti naglasov za posamezni zlog glede na število zlogov v besedi (Gros, 1997; Šef, 1998). V ta sklop lahko štejemo tudi samodejno analizo naglašenosti večzložnih besed, ki jo je opravila E. Tičević (2000). Avtorica je na osnovi slovarja izgovarjav s 34.880 besedami, do katerega je prišla z avtomatskim naglaševanjem in kasnejšo ročno obdelavo (če je bilo za katero besedno obliko možnih več različnih izgovarjav, se je odločila za tisto, ki po njenem mnenju nastopa najpogosteje), izpisala vse nastopajoče strukture zlogov v besedah ter poiskala najbolj verjetno naglasno mesto za posamezno strukturo. Poleg tega je določila verjetnosti pojavitve širokih oziroma ozkih samoglasnikov po strukturah.

2. Motivacija

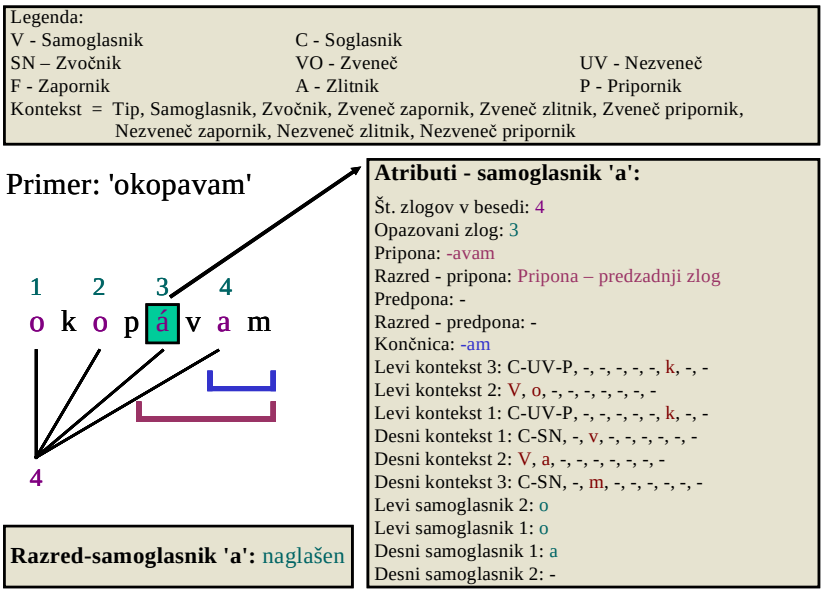
V sistemu GOVOREC naglaševanje besed v osnovi temelji na slovarsko podprti analizi besedila (Šef 2001). Za vse besede, ki so v slovarju, prepišemo mesto in vrsto tako dinamičnega kot tonemskega naglasa (ob upoštevanju oblikoslovnih podatkov in pravil za obravnavo homografov). Vendar pa še tako obsežen slovar ne more pokriti redkih in novonastalih besed. Izkazalo se je, da preprosti modeli naglaševanja nepoznanih besed ne dajo želenih rezultatov.

Ljudje lahko (pogosto) izgovorimo besede pravilno, čeprav jih nismo še nikdar slišali. To sposobnost želimo zajeti z avtomatskimi postopki učenja, ki omogočajo doseganje boljših rezultatov. Učinkovita pravila tudi zmanjšajo obseg slovarja (ta vsebuje le še besede, ki jih pravila ne pokrivajo), s čimer se zmanjšajo potrebe po pomnilniku. To je zlasti pomembno pri uporabi tovrstnih programov v dlančnikih, mobilnih telefonih, govorečih slovarjih, itd.

3. Metodologija

Razvili smo dvostopenjski model naglaševanja nepoznanih slovenskih besed. Za vsak samoglasnik in soglasnik 'r' v besedi najprej pogledamo, če je naglašen; za naglašene glasove določimo tudi tip naglasa. To naredimo s pomočjo odločitvenih dreves, ki smo jih dobili z metodami strojnega učenja (upoštevamo 66 atributov). Uporabljen je bil Quinlanov algoritem See5/C5.0 (Rulequest Research, 2002). Sledi popravljanje tako dobljenih rezultatov glede na število naglasov v besedi in dolžino besede. Pri najbolj preprosti različici algoritma v primeru večih naglasov naključno izberemo enega samega. Kadar je neka beseda napačno naglašena, se napaka ponavadi nahaja na dveh zlogih: na zlogu, ki bi moral biti naglašen (pa ni) in na zlogu, ki je naglašen (pa ne bi smel biti).

Zgenerirali smo domeno s primeri. To smo nato razdelili na šest poddomen; po eno za vsak samoglasnik in soglasnik 'r'. Za vsako poddomeno smo zgradili ločen model (odločitveno drevo). Osredotočili smo se tako na učinkovitost zgrajenih modelov, kakor tudi na interpretabilnost dobljenih rezultatov. Za odločitvena drevesa smo se odločili, ker jih lahko zlahka pretvorimo v pravila. Učna in testna množica sta bili ločeni. Rezultate poizkusov smo izračunali na nivoju zlogov in na nivoju besed.



Slika 1. Atributi za tretji samoglasnik ('o') za slovensko besedo "*okopavam*"

4. 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.

4.1. Pridobivanje podatkov

Slovar slovenskega knjižnega jezika (SSKJ) vsebuje le besede v njihovih osnovnih oblikah, zato smo bili prisiljeni zgraditi nov fonetični slovar v elektronski obliki. Ta vsebuje 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. Večina dela je bila opravljenega avtomatično z uporabo v ta namen razvitega morfološkega analizatorja (ca. 50.000 vrstic programske kode v C-ju) 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 192.132 besed. Pri tem smo odstranili večkratne ponovitve posamezne besedne oblike z enako izgovarjavo, a različno morfološko oznako. Kot rezultat smo dobili 700.340 zlogov (samoglasnikov). Te smo razdelili na učno in testno množico. Učna množica je vsebovala 140.821 besed (513.309 samoglasnikov), testna množica pa 51.311 besed (187.031 samoglasnikov). Pri tem so besede (osnovne oblike in izpeljanke) v testni množici pripadale različnim lemam kot besede v učni množici. Tako si učna in testna množica nista bili preveč podobni. Ker pa so nepoznane besede pogosto izpeljanke obstoječih besed v slovarju izgovarjav, so rezultati na dejanskih podatkih (nepoznane besede v besedilu, ki se sintetizira) po vsej verjetnosti celo nekoliko boljši od prikazanih. Tej tezi v prid gre tudi dejstvo, da imajo nepoznane besede (za razliko od najpogosteje uporabljanih) bolj standarden način izgovarjave, z brez oz. manj izjemami.

4.2. Opis podatkov

Učna in testna množica sta bili razdeljeni v več podmnožic glede na samoglasnike in soglasnik 'r'. Tako smo dobili šest ločenih učnih problemov. Število primerov v tako dobljenih podmnožicah prikazuje Tabela 1. Distribucije razredov v učni in testni množici so skoraj enake, nekaj odstopanja je zaznati le pri soglasniku 'r'.

Tabela 1: Število primerov v učni in testni množici

	A	E	I	O	U	R
Učna množica	142041	119227	116486	100295	28104	7156
Testna množica	50505	47169	41156	35513	9870	2818

Vsak primer je opisan s 66 atributi, vključno z razredom, ki predstavlja tip dinamičnega naglasa. Njegove vrednosti so 'Nenaglašen', 'Naglašen-širok', 'Naglašen-ozek', 'Nenaglašen-polglasnik' in 'Naglašen-Polglasnik'. Dejavniki, ki ustrezajo preostalim 65 atributom so:

- število zlogov v besedi (1 atribut),
- položaj opazovanega samoglasnika (zloga) v besedi (1 atribut),
- prisotnost predpone oz. pripone v besedi in razreda, ki mu pripada (4 atributi),
- končnica besede (1 atribut),
- levi in desni kontekst opazovanega samoglasnika (tip in ime grafema za tri znake levo in desno, dva samoglasnika levo in desno od opazovanega samoglasnika) (58 atributov).

Dosedanje metode učenja izgovarjave nepoznanih besed temeljijo na predpostavki, da se vsa potrebna informacija v celoti nahaja v nizu znakov, ki sestavljajo besedo. Pri slovenščini pa sta mesto in tip dinamičnega naglasa odvisna še od morfoloških karakteristik besede. Za pravilno izgovarjavo besede tako potrebujemo še njene oblikoslovne podatke. Naš sintetizator govora vsebuje oblikoslovni označevalnik, ki je sposoben obravnavati nepoznane besede, vendar trenutno še ni dovolj zanesljiv (zaradi premalo obsežnega morfološko označenega korpusa besedil). Drugi razlog, da te informacije nismo vključili v naš model (čeprav je preprosto izvedljivo in v prihodnosti to tudi nameravamo storiti), je potreba po zmanjšanju obsega slovarja izgovarjav za uporabo v dlančnikih. Poleg tega je del morfološke informacije že vsebovan v predponah, priponah in besednih končnicah.

Ugotovili smo, da dobimo boljše rezultate in bolj strnjena odločitvena drevesa, če predstavimo kontekst opazovanega samoglasnika s tipom grafema (samoglasnik ali soglasnik, tip soglasnika) in ne z imenom samega grafema. Ime grafema je označeno posebej v enem od atributov, ki pojasnjujejo posamezne tipe grafemov (npr. atribut 'Samoglasnik' lahko vsebuje naslednje vrednosti: 'a', 'e', 'i', 'o', 'u', '-' (ni samoglasnik)). Primer prikazuje Slika 1.

5. Eksperiment

Na šestih domenah, ki ustrezajo petim samoglasnikom in soglasniku 'r', smo učili odločitvena drevesa (DT in boosted DT), kot je to implementirano v See5 sistemu (Rulequest Research. 2002; Quinlan 1986). Vrednotenje rezultatov smo opravili na ločeni testni množici.

Kot parameter za rezanje dreves smo uporabili minimalno število primerov v listih. Oba načina učenja odločitvenih dreves (DT in boosted DT) smo primerjali med seboj in sicer za parametre rezanja dreves med 2 in 1000 minimalnimi primeri v listih. Rezultati so prikazani v Tabeli 2 ter na Sliki 2, Sliki 3 in Sliki 4. Napaka je bila najmanjša pri drugem načinu učenja (boosted DT) ob minimalnem rezanju dreves (minimalno 2 primera v listih).

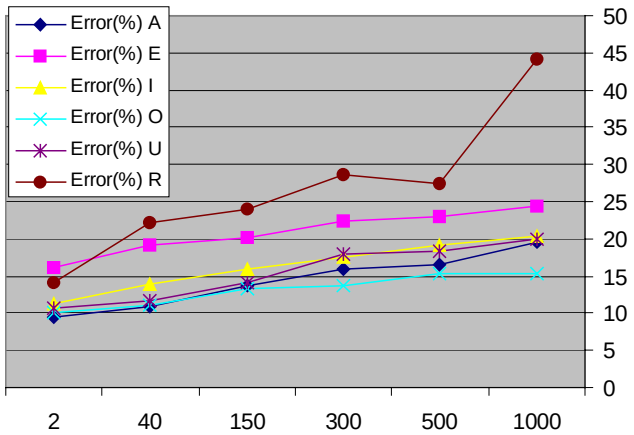
Tabela 2: Napaka pri različnih metodah učenja (DT in boosted DT) in pri različnih parametrih rezanja dreves

	Odločitvena drevesa – min. št. primerov v listih							
	1000		500		300		150	
	DT	B DT	DT	B DT	DT	B DT	DT	B DT
A	19.6	13.9	16.5	10.6	16	9.8	13.7	8.6
E	24.4	21.8	22.9	19.5	22.3	16.8	20.2	14.3
I	20.3	16.4	19.2	14.4	17.6	13.0	15.9	11.4
O	15.3	13.4	15.4	12.3	13.7	11.4	13.3	10.3
U	20.0	12.8	18.3	12.8	18.0	12.1	14.1	9.7
R	44.1	28.2	27.5	23.0	28.6	24.4	20.5	23.9
Zlog	20.5	16.5	18.7	14.3	17.8	12.9	15.8	11.3
Beseda	37.4	30.1	34.2	26.0	32.4	23.5	28.8	20.5

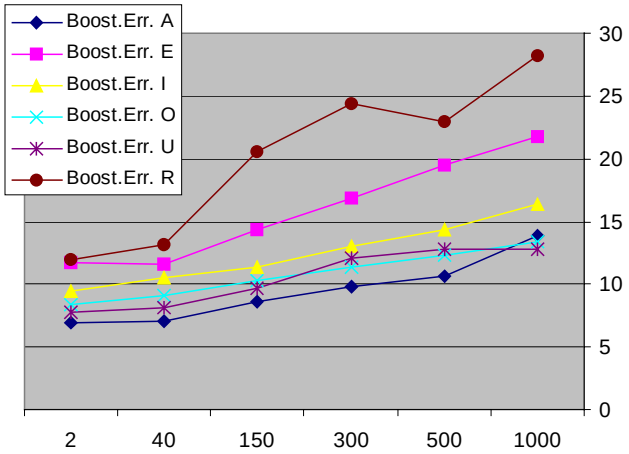
	Odločitvena drevesa				Pravila za
	40		2		
	DT	B DT	DT	B DT	
A	10.9	7.1	9.5	6.9	22.8
E	19.1	11.6	16.1	11.7	29.4
I	13.9	10.5	11.3	9.5	22.7
O	11.1	9.1	10.0	8.4	24.0
U	11.7	8.1	10.6	7.8	22.9
R	22.1	13.1	14.2	11.9	33.6
Zlog	13.9	9.5	11.8	9.1	24.7
Beseda	25.3	17.3	21.5	16.5	47.9

Kot smo pričakovali, se avtomatske metode učenja z uporabo odločitvenih dreves odrežejo veliko bolje kot obstoječa pravila (Šef, 1998; Toporišič, 1984) za naglaševanje. Napaka se v najboljšem primeru (boosted DT, minimalno 2 primera v listih) zmanjša za 31,4 %. Celo pri uporabi najbolj porezanih dreves je napaka manjša kot ob uporabi ročno dobljenih pravil.

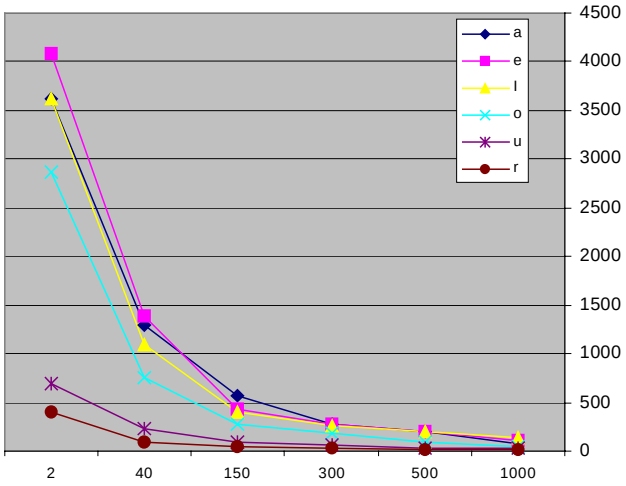
Glede interpretabilnosti modela lahko ugotovimo nekaj značilnosti odločitvenih dreves: (1) so zelo široka, (2) so globoka, (3) koren drevesa ostaja ob naraščanju velikosti drevesa nespremenjen. Glede na (1) in (2) lahko zaključimo, da je tako dobljena drevesa težko interpretirati oz. da ne obstajajo neka preprosta pravila naglaševanja.



Slika 2. Napaka pri prvi metodi učenja (DT) ob različnih parametrih rezanja dreves



Slika 3. Napaka pri drugi metodi učenja (boosted DT) ob različnih parametrih rezanja dreves



Slika 4. Velikost odločitvenih dreves (DT) iz Tabele 2

6. Zaključek

Predstavili smo dvostopenjski model naglaševanja nepoznanih slovenskih besed. Model temelji na metodah strojnega učenja ob uporabi odločitvenih dreves in je sposoben pravilno naglasiti nepoznano slovensko besedo v več kot 83 % primerov. Rezultati so znatno boljši od do sedaj uporabljenih ročno pridobljenih pravil (52 % natančnost).

Eksperimenti so potrdili že v uvodu omenjeno tezo, da za naglaševanje slovenskih besed ne obstajajo neka preprosta pravila.

7. Literatura

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

J. Gros. 1997. *Samodejno tvorjenje govora iz besedil*. Doktorsko delo. Fakulteta za elektrotehniko, Univerza v Ljubljani.

J. R. Quinlan. 1986. Induction of Decision Tress, *Machine Learning*, 1:81-106

Rulequest Research. 2002. *See5 system*. (http://www.rulequest.com/see5-info.html).

T. Šef. 1998. *Sistem za govorno posredovanje obvestil o prostih delovnih mestih*. Magistrsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

T. Šef. 2001. *Analiza besedila v postopku sinteze slovenskega govora*. Doktorsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

E. Tičević. 2000. *Samodejna analiza naglašenosti večzložnih besed*. Diplomsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani.

J. Toporišič. 1984. *Slovenska slovnica*. Založba Obzorja, Maribor.