Elektrotehniški vestnik 73(2-3): 99-104, 2006
Electrotechnical Review: Ljubljana, Slovenija

Avtomatsko naglaševanje nepoznanih besed pri sintezi slovenskega govora

Tomaž Šef

Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana, Slovenija
E-pošta: tomaz.sef@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, ali 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 leksikona MULTEXT-East, 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 kompleksen problem, ki ga z relativno preprostimi pravili ni mogoče učinkovito rešiti.

Ključne besede: sinteza govora, naravni jezik, naglaševanje besed

Automatic Lexical Stress Assignment of Unknown Words for the Highly Inflected Slovenian Language

Extended abstract. Grapheme-to-phoneme conversion is an essential task in any text-to-speech system. It can be described as a function mapping the spelling form of words to a string of phonetic symbols representing the pronunciation of the word. An important task when building rule-based grapheme-to-phoneme transcription is dealing with words which are not encapsulated in vocabularies.

It is well known that the correspondence between spelling and pronunciation can be rather complicated. Usually it involves stress assignment and letter-to-phone transcription. Unlike the majority of world languages, it is straightforward to convert the word into its phonetic representation in the Slovenian language, when the stress type and position are known. It can be done on the basis of less than 100 context-dependent letter-to-sound rules (composed by well-versed linguists) with the accuracy of over 99%. A crucial problem is the determination of the lexical stress type and position. A lexical stress in the Slovenian language can be located almost arbitrarily on any syllable in the word, therefore it is often assumed to be "unpredictable".

Although humans pronounce very well even words they have never heard before, human experts have not been able to synthesise accurate rules for accentuation in phonetically complex languages. We performed several experiments with different machine-learning and data-mining tools to create accentuation rules for the Slovenian language. We wanted to find out if it is possible to design better accentuation rules than the ones defined by the experts, and to test if these rules are complex or compact.

A two-step pronunciation model was applied. In the first step we used a machine-learning model (DT or boosted DT) to predict the lexical stress on a vowel and consonant “r”. In the second step the lexical stress of the whole word was predicted. If the DT model in the first step predicted more than one stressed vowel, one of them was randomly chosen. If the prediction of the lexical stress of a whole word was false, then typically two incorrect lexical stresses were made: one on the right syllable, which was not stressed, and the other on the syllable incorrectly predicted to be stressed.

The MULTEXT-East Slovenian Lexicon was used for data sets, supplemented with lexical stress marks. For the first step we generated a domain consisting of vowels and consonant “r”. The domain was separated into six sub-domains, one for each vowel and consonant “r”. For each vowel and consonant 'r' we trained a separate model (DT and boosted DT) on the learning set and evaluated it on the corresponding test set. The error was then calculated on the level of a syllable and a word.

Prejet 2. november, 2005
Odobren 3. maj, 2006

The accuracy achieved by decision trees significantly surpassed all previous results. However, the sizes of the trees indicate that accentuation in Slovene is a very complex problem. We were not able to find a simple solution in the form of relatively compact rules using several ML and DM systems. This also indicates that creating accurate accentuation rules is a task that might require artificial intelligence.

Key words: text-to-speech synthesis, natural language, lexical stress assignment.

1 Uvod

Mesto naglasa je zlog, na katerem ima beseda jakostno ali tonemsko izrazitost. V nasprotju z nekaterimi drugimi jeziki (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 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 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) [1] in statistikah, ki so podajale verjetnost naglasov za posamezni zlog glede na število zlogov v besedi [2, 3]. V ta sklop lahko štejemo tudi samodejno analizo naglašenosti večzložnih besed, ki jo je opravila E. Tičević [4]. Avtorica je na podlagi slovarja izgovarjav s 34.880 besedami, do katerega je prišla z avtomatskim naglaševanjem in poznejšo ročno obdelavo (če je bilo za katero besedno obliko mogočih 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 [5]. 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 novih 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 vključujejo), s čimer se zmanjšajo potrebe po pomnilniku. To je pomembno zlasti 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, ali 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 [6]. Sledi popravljanje tako dobljenih rezultatov glede na število naglasov v besedi in dolžino besede. Pri najpreprostejši različici algoritma v primeru večih naglasov naključno izberemo enega samega. Kadar je neka beseda napačno naglašena, je napaka ponavadi 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 ravni zlogov in na ravni besed.

4 Podatki

Prvi pogoj 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.

Tabela 1: Število primerov v učni in testni množici
Table 1: Number of examples in learning and test sets

		A	E	I	O	U	R
Učna množica	Vsi	142.041	119.227	116.486	100.295	28.104	7.156
	Nenaglašeni	70,05 %	70,17 %	68,31 %	78,67 %	62,75 %	-
	Naglašeni (široki)	29,95%	8,26 %	31,48 %	3,91 %	37,25 %	-
	Naglašeni (ozki)	-	20,13%	-	17,41 %	-	-
	Nenaglašeni polglasniki	-	1,36 %	-	-	-	52,21 %
	Naglašeni polglasniki	-	0,08 %	-	-	-	47,21 %
Testna množica	Vsi	50.505	47.169	41.156	35.513	9.870	2.818
	Nenaglašeni	71,52%	67,76 %	65,07 %	79,98	61,58 %	-
	Naglašeni (široki)	28,48 %	9,39 %	34,93 %	3,73	38,42 %	-
	Naglašeni (ozki)	-	21,29 %	-	16,29	-	-
	Nenaglašeni polglasniki	-	1,53 %	-	-	-	40,81 %
	Naglašeni polglasniki	-	0,04 %	-	-	-	59,19 %

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 leksikon MULTEXT-East [7], 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 (okrog 50.000 vrstic programske kode v C) in elektronske verzije SSKJ. Takšna določitev mesta naglasa je bila neuspešna v približno 0,2 odstotka primerih. Poleg tega je algoritem predlagal, da dodatno preverimo še nekaj manj kot en odstotek besed. V vseh teh primerih smo delo opravili ročno. Na koncu smo še enkrat pregledali celoten slovar.

Za graditev 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 (v nasprotju z najpogosteje uporabljanimi) bolj standarden način izgovarjave, z manj izjemami oziroma brez njih.

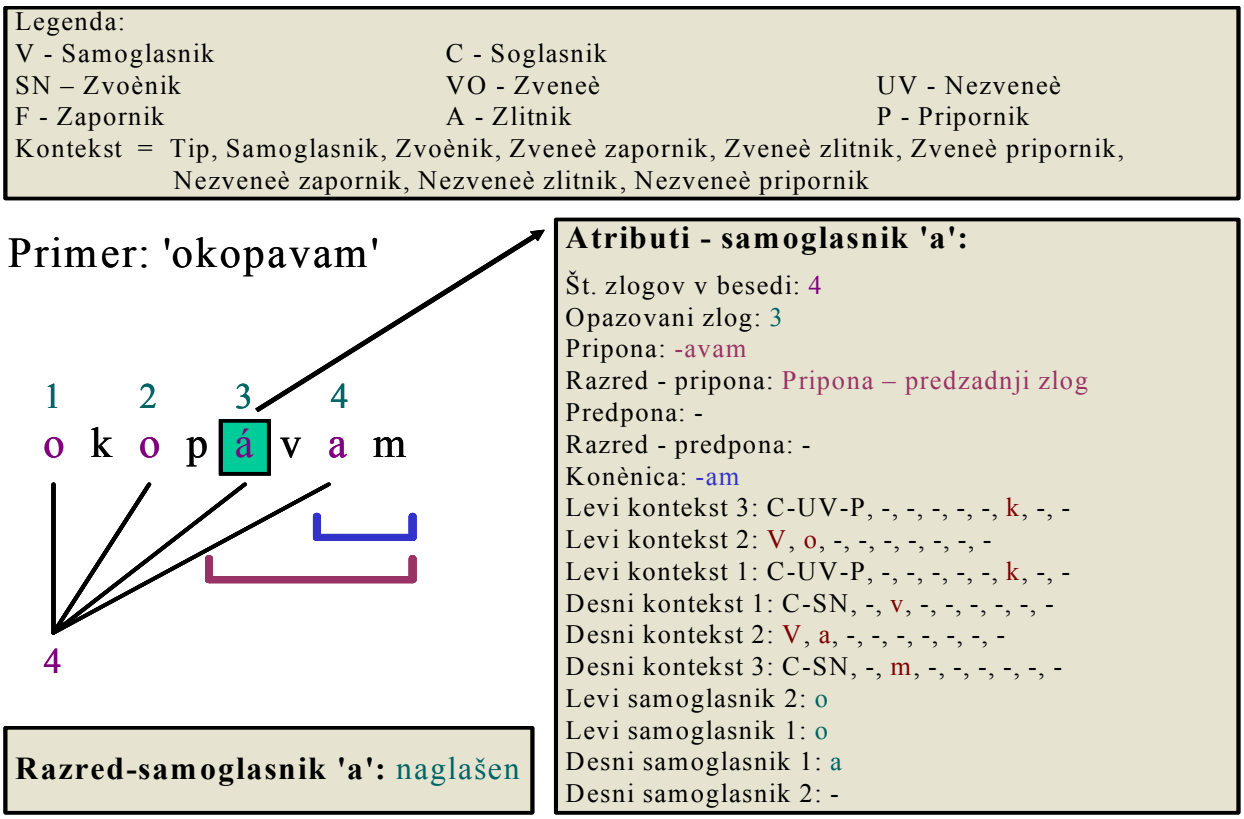
4.2 Opis podatkov

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'.

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).



Slika 1: Atributi za tretji samoglasnik ('a') za slovensko besedo "okopavam"
Figure 1: Attributes for the third vowel ('a') of the Slovenian word "okopavam" (engl. "I earth up")

Večina avtomatskih metod učenja izgovarjave nepoznanih besed temelji 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 Poskus

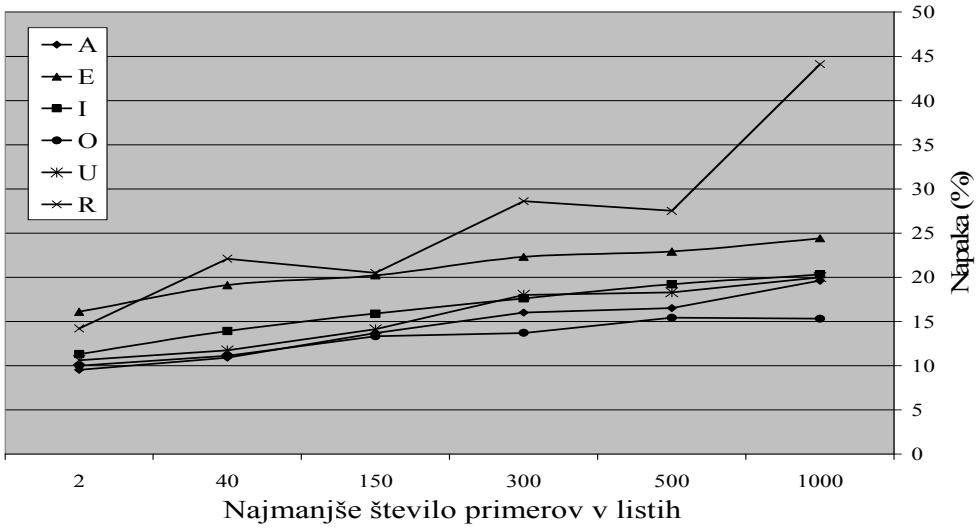
Na šestih domenah, ki ustrezajo petim samoglasnikom in soglasniku 'r', smo učili odločitvena drevesa (DT in boosted DT), kot je to implementirano v sistemu See5 [6, 8]. Vrednotenje rezultatov smo opravili na ločeni testni množici.

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 in 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).

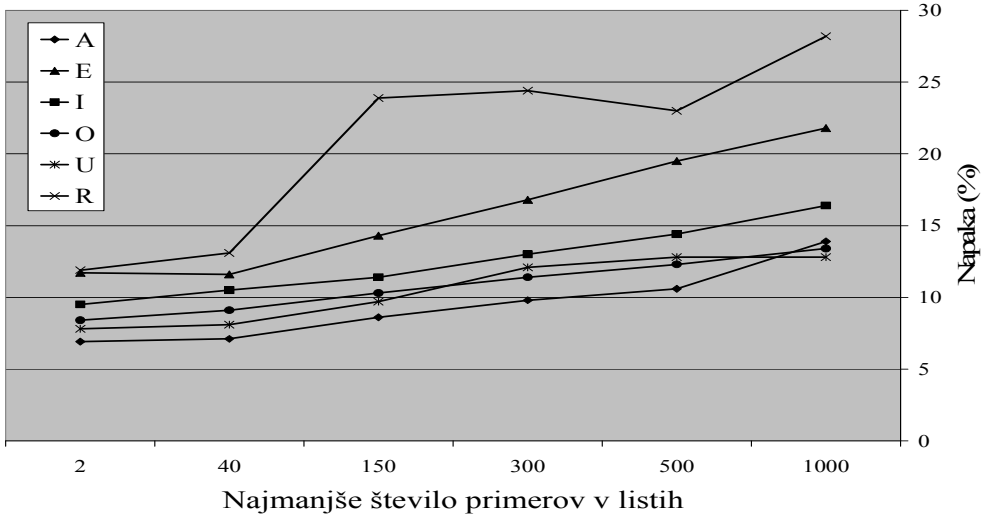
Kot smo pričakovali, se avtomatske metode učenja z uporabo odločitvenih dreves odrežejo veliko bolje kot obstoječa pravila [3, 1] za naglaševanje. Napaka se v najboljšem primeru (boosted DT, minimalno 2 primera v listih) zmanjša za 31,4 odstotka. Celo pri uporabi najbolj porezanih dreves je napaka manjša kot pri uporabi ročno dobljenih pravil.

Tabela 2: Napaka pri različnih metodah učenja (DT in boosted DT) in pri različnih parametrih rezanja dreves
Table 2: Error of DT classifier, Boosted DT classifier and grammatical rules on vowels, a syllable and a word

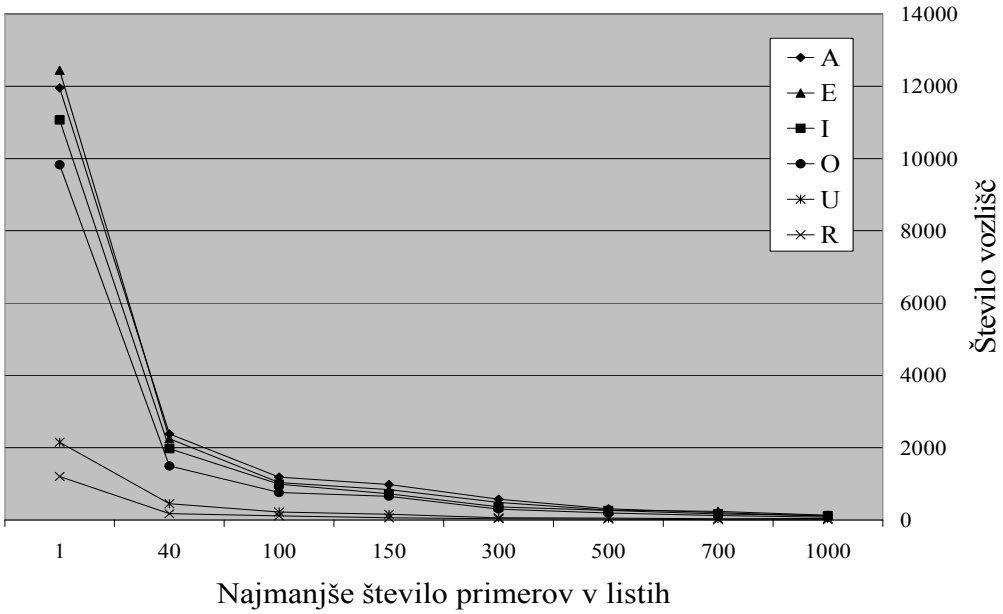
	Odločitvena drevesa – min. št. primerov v listih												Pravila za naglaševanje
	1000		500		300		150		40		2		
	DT	B DT	DT	B DT	DT	B DT	DT	B DT	DT	B DT	DT	B DT	
A	19,6	13,9	16,5	10,6	16	9,8	13,7	8,6	10,9	7,1	9,5	6,9	22,8
E	24,4	21,8	22,9	19,5	22,3	16,8	20,2	14,3	19,1	11,6	16,1	11,7	29,4
I	20,3	16,4	19,2	14,4	17,6	13,0	15,9	11,4	13,9	10,5	11,3	9,5	22,7
O	15,3	13,4	15,4	12,3	13,7	11,4	13,3	10,3	11,1	9,1	10,0	8,4	24,0
U	20,0	12,8	18,3	12,8	18,0	12,1	14,1	9,7	11,7	8,1	10,6	7,8	22,9
R	44,1	28,2	27,5	23,0	28,6	24,4	20,5	23,9	22,1	13,1	14,2	11,9	33,6
Zlog	20,5	16,5	18,7	14,3	17,8	12,9	15,8	11,3	13,9	9,5	11,8	9,1	24,7
Beseda	37,4	30,1	34,2	26,0	32,4	23,5	28,8	20,5	25,3	17,3	21,5	16,5	47,9



Slika 2: Napaka pri prvi metodi učenja (DT) ob različnih parametrih rezanja dreves
Figure 2: DT error under different pruning



Slika 3: Napaka pri drugi metodi učenja (boosted DT) ob različnih parametrih rezanja dreves
Figure 3: Boosted DT error depending on different pruning values



Slika 4: Velikost odločitvenih dreves (DT)
Figure 4: Size of the DT learned on the whole set

Glede interpretabilnosti modela 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 zapišemo, da je tako dobljena drevesa težko interpretirati oz. da ni nekih preprostih pravil naglaševanja [9].

6 Sklep

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

Poskusi so potrdili že v uvodu omenjeno tezo, da za naglaševanje slovenskih besed ni preprostih pravil.

7 Literatura

[1] J. Toporišič, *Slovenska slovnica*, Založba Obzorja, Maribor, 1984.

[2] J. Gros, Samodejno tvorjenje govora iz besedil, *doktorska disertacija*, Fakulteta za elektrotehniko, Univerza v Ljubljani, 1997.

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

[4] E. Tičević, Samodejna analiza naglašenosti večzložnih besed, *diplomska naloga*, Fakulteta za elektrotehniko, Univerza v Ljubljani, 2000.

[5] T. Šef, Analiza besedila v postopku sinteze slovenskega govora, *doktorska disertacija*, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2001.

[6] See5 system (*http://www.rulequest.com/see5-info. html*)

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

[8] J. R. Quinlan, Induction of Decision Tress, *Machine Learning 1*, str. 81-106, 1986.

[9] T. Šef, M. Gams, Data Mining for Creating Accentuation Rules, *Applied Artificial Intelligence*, Taylor & Francis, 18:395-410, 2004.

Tomaž Šef je diplomiral leta 1995 na Fakulteti za elektrotehniko in računalništvo v Ljubljani. Leta 1998 je magistriral, leta 2001 pa doktoriral na Fakulteti za računalništvo in informatiko v Ljubljani. Zaposlen je na Odseku za inteligentne sisteme na Institutu "Jožef Stefan", kjer se ukvarja z umetno inteligenco, komunikacijo človek-stroj, procesiranjem naravnega jezika, sintezo govora ter zajemanjem tekstovnih in govornih baz podatkov.