Uporaba oblikoslovnih oznak pri naglaševanju 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

Accentuation of Slovenian Words using Morphological Tags

Stress assignment within words is a basic task in 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 Slovenian language. Therefore, we use machine learning techniques to create accentuation rules for the Slovenian language. In this article we compare the classification accuracy of accentuation when using different sets of attributes. We experiment with attributes, that define only the context of the observed grapheme, attributes, that consider also the basic properties of the word (such as the number of syllables, prefix, suffix, etc.), and attributes, that take into account also the morphological information of the word. As shown in the results, the best classification accuracy is achieved when all attributes are considered. This suggests that the morphological information helps 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 težka 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 izgovorjavo 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 izgovorjave nepoznanih slovenskih besed temeljijo na predpostavki, da se vsa potrebna informacija v celoti nahaja v nizu znakov, ki sestavljajo besedo [7, 2]. Ker sta pri slovenščini mesto in tip dinamičnega naglasa odvisna tudi od oblikoslovnih karakteristik besede, za pravilno izgovorjavo besede potrebujemo tudi to informacijo. V pričujočem članku prvič 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. Rezultate primerjamo z rezultati naglaševanja, ki jih dobimo, če teh informacij ne uporabimo, in z rezultati naglaševanja v primeru, ko upoštevamo le kontekst opazovanega grafema.

V naslednjem razdelku opisujemo podatke, ki smo jih imeli na voljo za poskuse ter način pridobivanja leteh. 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 izgovorjave 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 elektron-

ski obliki. Ta vsebuje okoli 600.000 besednih oblik, kar ustreza 20.000 lemam. Kot osnovo smo uporabili MULTEXT-East leksikon [3], 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 (cca. 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 izgovorjavo, a različno oblikoslovno oznako. Kot rezultat smo dobili 2.017.858 zlogov. Te smo razdelili na tri učne in testne množice, da smo lahko uporabili metodo trikratnega prečnega preverjanja. Vse besede z isto lemo smo vedno postavili samo v eno množico (bodisi učno bodisi testno) in tako poskrbeli, da niso bile testne množice preveč podobne učnim.

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

2.2. Določitev atributov

Za strojno učenje naglaševanja grafemov smo imeli na voljo 74 atributov, ki so prikazani na sliki 1:

- število zlogov v besedi (1 atribut);
- položaj 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);

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

- tip naglasa na izbranem samoglasniku (1 atribut) – to je razred, ki ga napovedujemo.

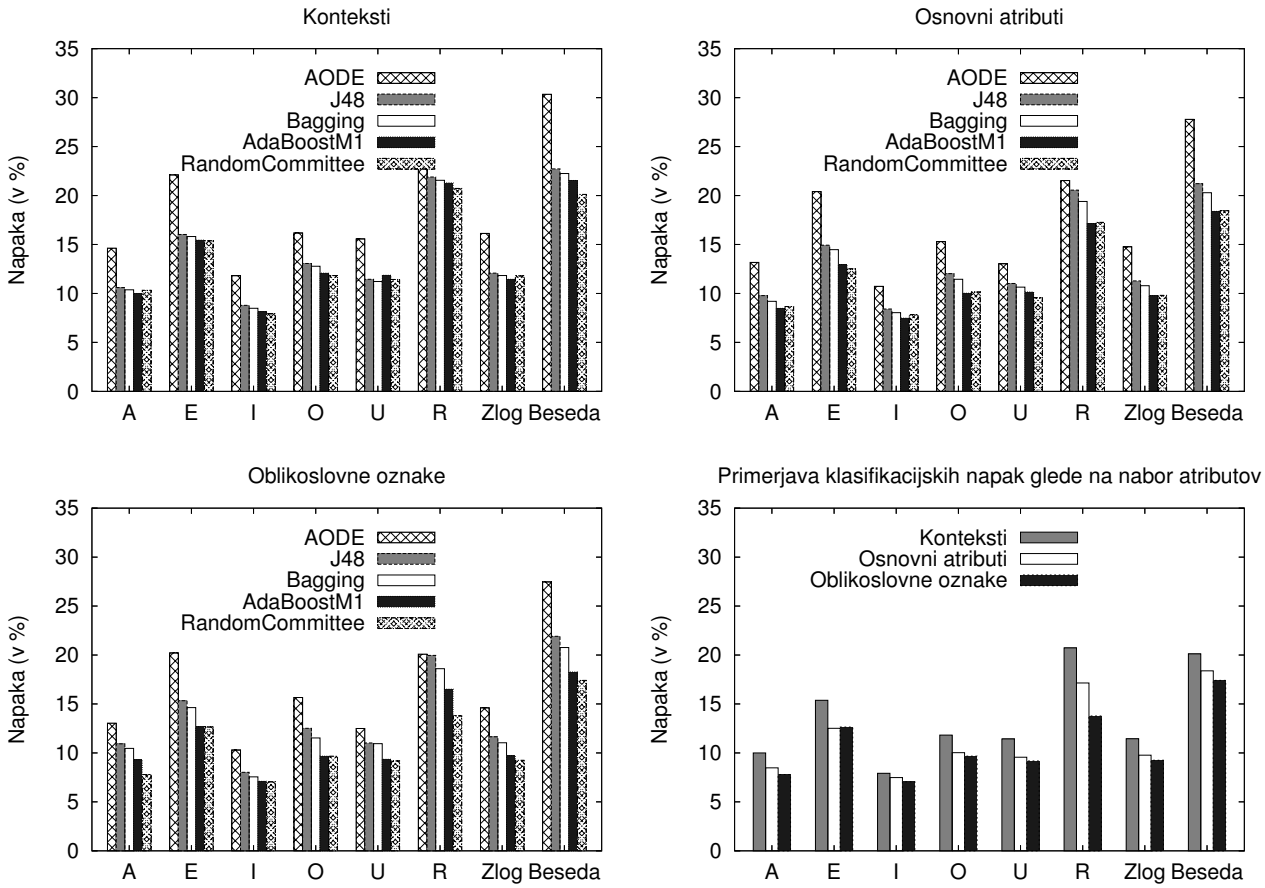
Želeli smo ugotoviti, s katerimi atributi dobimo najboljše rezultate. Zato smo poskuse opravili s tremi različnimi nabori atributov:

1. *Kontekst:* izbrali smo 54 atributov, ki določajo kontekst opazovanega grafema (na sliki 1 so prikazani pod številko 1) in atribut naglas.
2. *Osnovni atributi:* poleg atributov, ki opisujejo kontekst, smo upoštevali še osnovne atribute in tako dobili 67 atributov (to so atributi 1 in 2 na sliki 1).
3. *Oblikoslovne oznake:* vzeli smo vseh 74 atributov, vključno s tistimi, ki vsebujejo oblikoslovne oznake opazovane besede.

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 [10]. 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. To so naslednje metode:



Slika 2: Rezultati izbranih metod strojnega učenja za tri nabore atributov: *konteksti*, *osnovni atributi* in *oblikoslovne oznake* ter primerjava najboljših rezultatov za tri nabore atributov

- *NaiveBayes* je metoda, pri kateri uporabljamo naivni Bayesov klasifikator – predvidevamo, da so vsi atributi med sabo neodvisni [6].
- *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 [9].
- *ADTree* generira alternirajoče odločitveno drevo [4].
- *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 [8].
- *Bagging* je meta metoda, ki generira t osnovnih modelov za klasifikacijo in nato posamezne primere klasificira glede na večinsko mnenje osnovnih modelov [1].
- *AdaBoostM1* je meta metoda, ki implementira algoritem boosting. Deluje podobno kot metoda bagging, le da vsak od t osnovnih modelov za klasifikacijo dopolnjuje prejšnjega [5].
- *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.

	A	E	I	O	U	R	Zlog	Beseda
Konteksti	10.00 %	15.38 %	7.92 %	11.82 %	11.43 %	20.73 %	11.45 %	20.12 %
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 %

Tabela 1: Primerjava klasifikacijskih napak glede na nabor atributov

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 niso narisana. Zaradi enakega razloga smo iz slike izpustili tudi metodo NaiveBayes, ki se je odrezala slabše kot metoda AODE. Kot najboljša metoda strojnega učenja (pri vseh treh naborih parametrov), se je izkazala meta metoda RandomCommittee. Za njo le malo zaostaja metoda AdaBoostM1.

Kot lahko vidimo iz slike 2 (desno spodaj) in iz tabele 1, naglaševanje samo z uporabo informacije o kontekstu prinese 11,5 % napako v klasifikaciji zloga in 20 % napako v klasifikaciji besede. Boljše rezultate dobimo, če dodamo še osnovne atribute, ki zmanjšajo napako v klasifikaciji zloga na 9.8 % in napako v klasifikaciji besede na 18.4 %. K večji točnosti klasifikacije pripomorejo tudi atributi, ki opisujejo oblikoslovne oznake besed. Ti dosežejo 9.3 % napako v klasifikaciji zloga in 17.4 % napako v klasifikaciji besede.

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. S poskusi smo pokazali, da obravnavanje zgolj konteksta opazovanega grafema ne prinese dovolj dobrih rešitev. Boljše rezultate dobimo, če upoštevamo še druge atribute, kot so število zlogov v besedi, predpone in pripone ipd., še boljše pa, če atributom dodamo tudi oblikoslovne oznake besede.

Literatura

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

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

[3] T. Erjavec. The MULTEXT-east slovene lexicon. In *Zbornik sedme Elektrotehniške in računalniške konference ERK'98*, volume B, pages 189–192, 1998.

[4] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceeding of the Sixteenth International Conference on Machine Learning*, pages 124–133, Bled, Slovenia, 1999.

[5] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc International Conference on Machine Learning*, pages 148–1556. Morgan Kaufmann, San Francisco, 1996.

[6] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intellingence*, pages 338–345. Morgan Kaufmann, San Mateo, 1995.

[7] M. Škrjanc, T. Šef, and M. Gams. Using decision trees for accentuation in the slovenian language. In *Proceedings STAIRS 2002, STarting Artificial Intelligence Researchers Symposium*, pages 135–144. Ios Press, Ohmsha, 2002.

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

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

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