Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Luka Krsnik

Napovedovanje naglasa slovenskih besed z metodami strojnega učenja

Mentor: prof. dr. Marko Robnik-Šikonja
Somentor: dr. Tomaž Šef

Ljubljana
21. 2. 2018

Vsebina

- predstavitev problema in pregled sorodnih del
- nevronske mreže
- arhitektura najboljših nevronskih mrež in njihove izboljšave
- ansambelska metoda
- rezultati
- uporabne rešitve

Predstavitev problema naglašanja

Naglas je zlog, na katerem je beseda jakostno ali tonemsko izrazita.

- naglasov besed ponavadi ne pišemo
- del problema pretvorbe grafema v fonem
- govorci se naučimo naglas skupaj z besedo in dobro naglašujemo nepoznane besede
- pomankljiva pravila za strojno naglaševanje

Cilji naloge

- reševanje problema z metodami strojnega učenja
- preizkus ansambelskih metod
- naglasitev prostodostopnega slovarja Sloleks
- izdelava aplikacije za naglaševanje povezanih besedil

Pregled sorodnih del

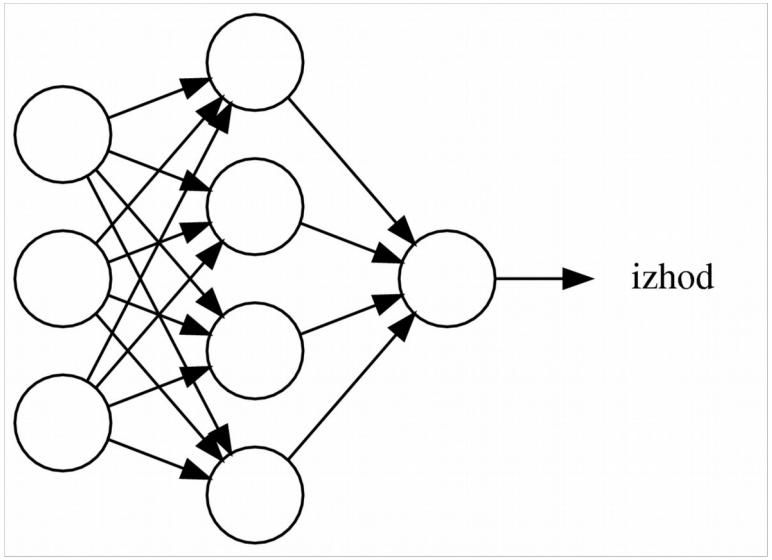
- metode strojnega učenja na problemu naglaševanja so boljše od pravil iz pravopisa
- preizkušeni markovski modeli in odločitvena drevesa
- dobri rezultati globokih nevronskih mrež na problemih procesiranja naravnega jezika

Nevronske mreže

- aktivacijska funkcija

$$z = \sum_j w_j x_j - b$$

$$izhod = \begin{cases} 0, \check{c}e\ velja\ z \leq 0 \\ 1, \check{c}e\ velja\ z > 0 \end{cases}$$

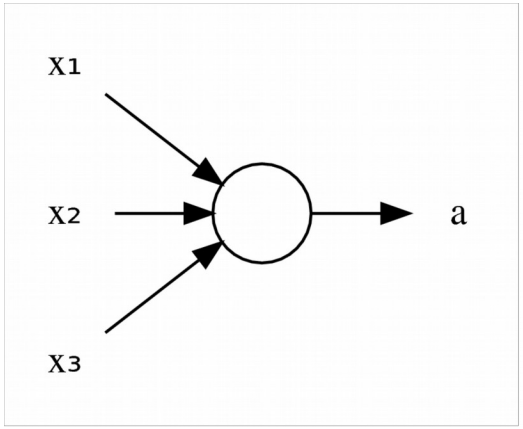


  - sigmoidna funkcija

$$\frac{1}{1+e^{-z}}$$

  - pragovna linearna funkcija ReLU

$$max(0, z)$$

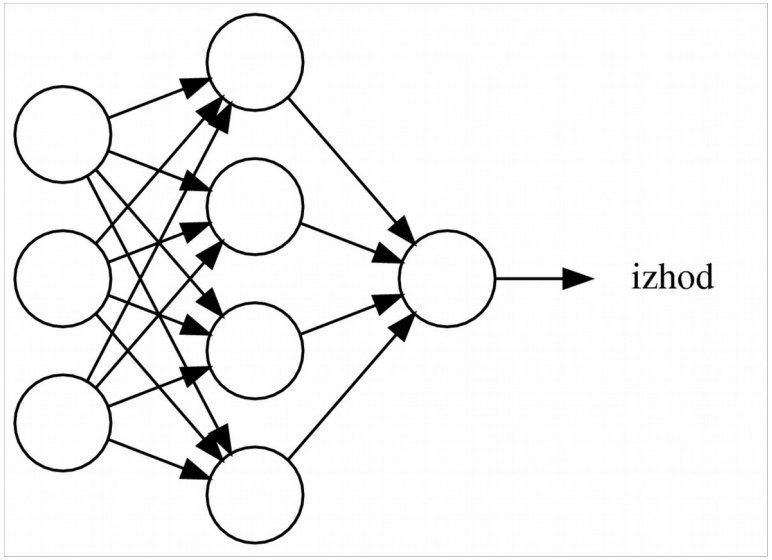


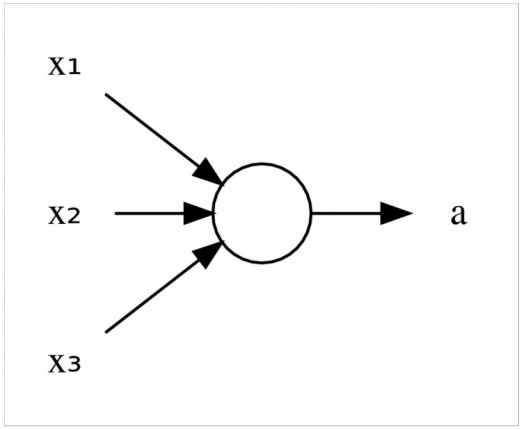
Nevronske mreže

- funkcija izgube
  - binarna križna entropija

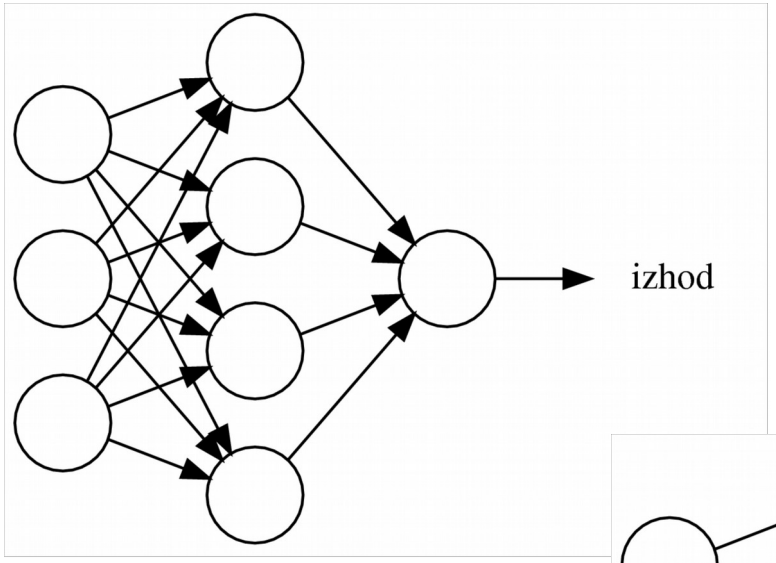
$$H(y, y') = -\sum_i y'_i \log(y_i) + (1 - y'_i) \log(1 - y_i)$$

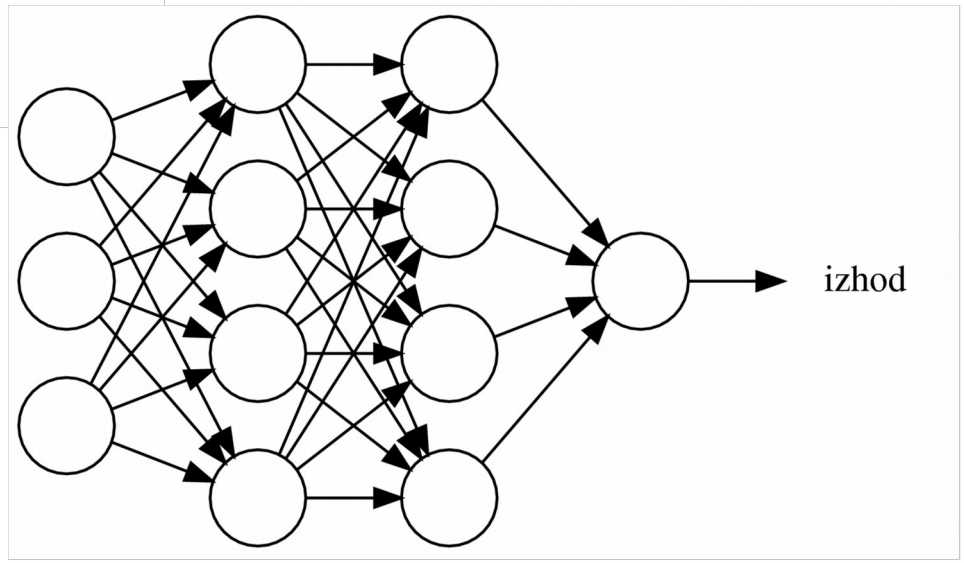
- gradientni spust





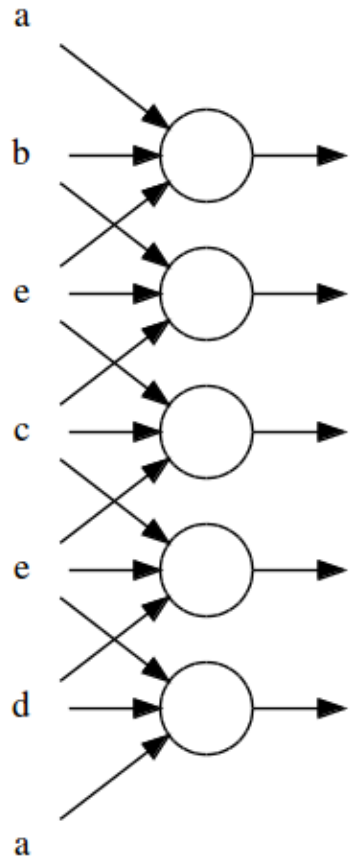
Globoke nevronske mreže





Konvolucijske nevronske mreže

- okno, ki se premika po besedi
- deljena utež pri implementaciji z mrežo
- več uteži na vsakem konvolucijskem nivoju



Podatkovna množica

- 540 000 besed
- 18 000 lem

Beseda	Lema	Morfološke informacije	Naglašena beseda
brati	brati	Vmn	bráti
berimo	brati	Vmmp1p	berímo
beri	brati	Vmmp2s	bêri

Krožnik - Ncmsn				
besedna vrsta	vrsta samostalnika	spol	število	sklon
N	c	m	s	n
samostalnik (Noun)	občno ime (common)	moški (male)	ednina (singular)	imenovalnik (nominative)

Delitev problema

- iskanje lokacije naglasa besede
- iskanje vrste dinamičnega naglasa
  - delitev naglasov po kvaliteti in kvantiteti

Simbol	Pomen
ä	kratek naglašeni a
ë	kratek naglašeni široki e
î	kratek naglašeni i
ö	kratek naglašeni široki o
ü	kratek naglašeni u
á	dolgi naglašeni a
é	dolgi naglašeni ozki e
ě	dolgi naglašeni široki e
í	dolgi naglašeni i
ó	dolgi naglašeni ozki o
ô	dolgi naglašeni široki o
ú	dolgi naglašeni u
ŕ	dolgi naglašeni r

Predstavitev podatkov mreži za iskanje mesta naglasa

- predstavitev vhodnih besed s črkami (primer ***abeceda***)
- predstavitev izhoda preko potencialnih mest naglasa (primer ***zamŕzniti***)

Črke vhodne besede	Indeksi črk
a	1
b	2
e	6
c	3
e	6
d	5
a	1

Seznam vseh črk		a	b	c	č	d	e	...
Kodiran zapis črke *c*	0	0	0	1	0	0	0	0

Potencialna mesta naglasa besede *zamrzniti*	a	r	i	i	...
Kodiran zapis mesta naglasa besede *zamrzniti*	0	1	0	0	0

Predstavitev podatkov mreži za iskanje mesta naglasa

- predstavitev morfoloških podatkov besede »leden« - Agpmsnn
- MULTEXT-East

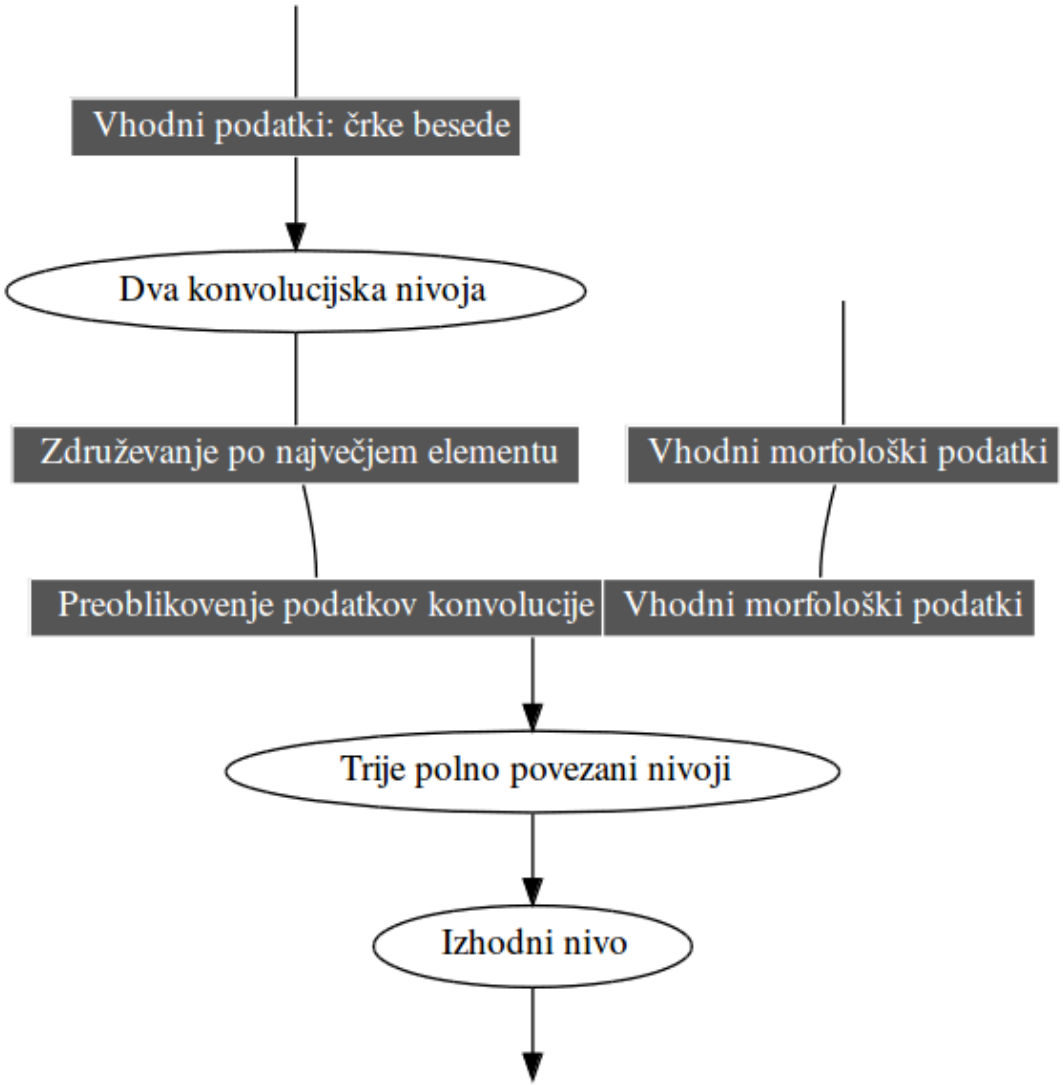
```
[ ['A',
   ['g', 's'],
   ['p', 'c', 's'],
   ['m', 'f', 'n'],
   ['s', 'd', 'p'],
   ['n', 'g', 'd', 'a', 'l', 'i'],
   ['-', 'n', 'y']
  ],
  ['C',
   ['c', 's']
  ],
  ...
]
```

```
[1,
 1, 0,
 1, 0, 0,
 1, 0, 0,
 1, 0, 0,
 1, 0, 0, 0, 0, 0,
 0, 1, 0,

 0,
 0, 0,


 ...
]
```

Arhitektura mreže za iskanje mesta naglasa

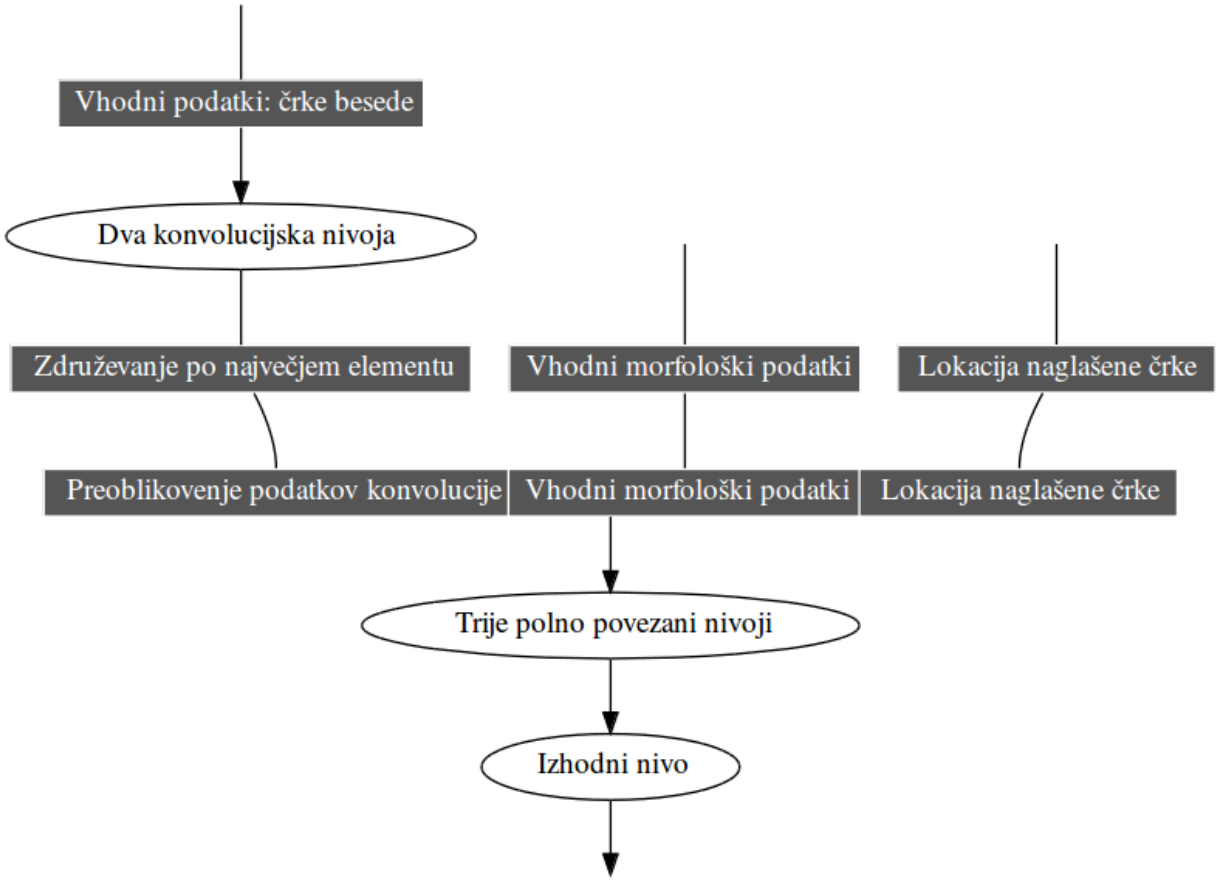


Predstavitev podatkov mreži za iskanje vrste naglasa

- enaka predstavitev vhodnih besed
- enaka predstavitev morfoloških podatkov
- podana lokacija mesta naglasa
- predstavitev izhoda preko vrst naglasa (primer ***zamŕzniti***)

Simbol	ä	ë	î	ö	ü	á	é	ě	í	ó	ô	ú	ŕ
Kodiran zapis izhoda	0	0	0	0	0	0	0	0	0	0	0	0	1

Arhitektura mreže iskanja vrste naglasa



Problem zlogovanja

- prostodostopnega zlogovalnika nismo našli
- open office in latex delilnik
- izdelava programskega zlogovalnika s pomočjo fiksnih pravil iz pravopisa

Predstavitev vhodov s približki zlogov in črkovanimi približki zlogov

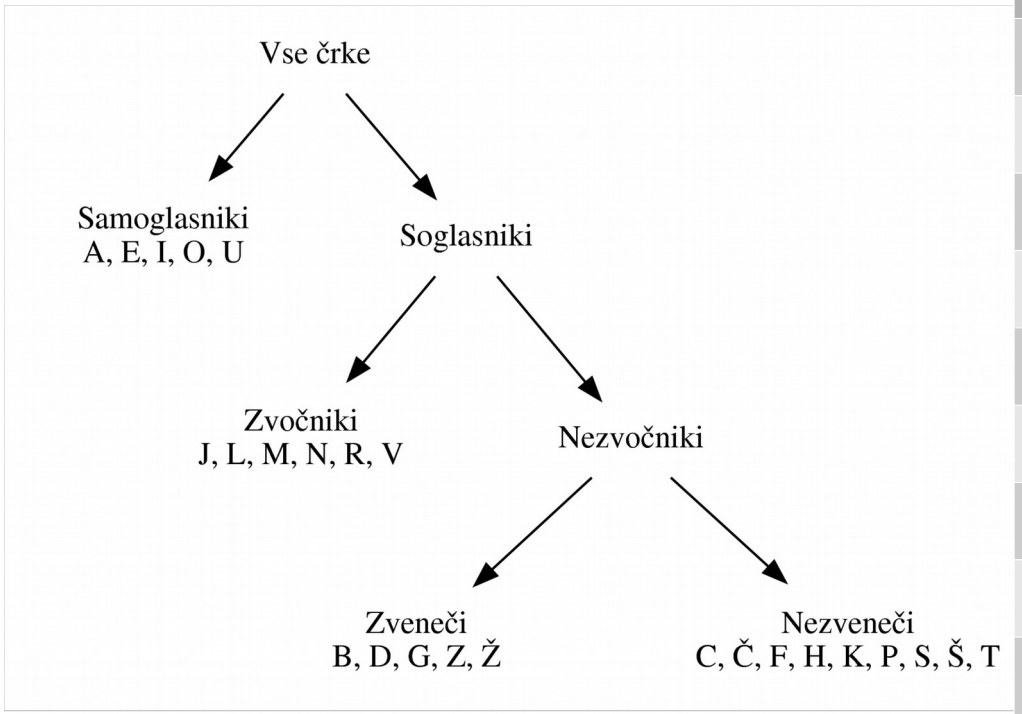
- približki zlogov
- črkovni približki zlogov

Približki zlogov besede	Indeksi zlogov
a	2
be	47
ce	310
da	407
...	0

Približki zlogov besede	Indeks prve črke	Indeks druge črke
a	1	0
be	2	6
ce	3	6
da	5	1
...	0	0

Izboljšave predprocesiranja črk

- črke podane skupaj z dodatnimi lastnostmi



	Kodiran zapis črke *c*
Brez črke	0
a	0
b	0
c	1
č	0
...	0
Samoglasnik	0
Soglasnik	1
Zvočnik	0
Nezvočnik	1
Zveneči	0
Nezveneči	1

Izboljšave predprocesiranja

- obrnjena beseda na vhodu in izhodu

Črke vhodne besede	Indeksi črk
a	1
d	5
e	6
c	3
e	6
b	2
a	1
...	0

Približki zlogov besede	Indeksi zlogov
da	407
ce	310
be	47
a	2
...	0

Potencialna mesta naglasa besede *zamrzniti*	i	i	r	a	...
Kodiran zapis mesta naglasa besede *zamrzniti*	0	0	1	0	0

Ansambelska metoda

- združevanje rezultatov večih napovednih modelov
- povprečenje
  - izračunamo povprečje vsake napovedi iz vseh napovednih modelov

Mesto naglasa	1.	2.	3.	...
1. napovedni model	0,8	0,2	0,0	...
2. napovedni model	0,6	0,4	0,6	...
3. napovedni model	0,4	0,6	0,6	...
Rezultat povprečenja	**0,6**	**0,4**	**0,4**	**...**
Napoved	**1**	**0**	**0**	**...**

Evalvacija

- delitev na tri množice:
  - učna (80 %)
  - validacijska (10 %)
  - testna (10 %)
- klasifikacijska točnost pravilno naglašenih celotnih besed

Končni rezultati

Napoved mesta naglasa	
Napovedni model	**Klasifikacijska točnost**
Mreža s črkovnim vhodom	88,54 %
Mreža s približkom zlogov	85,67 %
Mreža s črkovnim približkom zlogov	87,98 %
Neuteženo povprečenje zgornjih treh mrež	90,24 %

Napoved vrste naglasa	
Napovedni model	**Klasifikacijska točnost**
Mreža s črkovnim vhodom	96,06 %
Mreža s približkom zlogov	94,00 %
Mreža s črkovnim približkom zlogov	95,98 %
Neuteženo povprečenje zgornjih treh mrež	96,37 %

Napoved mesta in vrste naglasa	
Napovedni model	**Klasifikacijska točnost**
Mreža s črkovnim vhodom	85,65 %
Mreža s približkom zlogov	80,71 %
Mreža s črkovnim približkom zlogov	85,01 %
Neuteženo povprečenje zgornjih treh mrež	87,65 %

Primerjava rezultatov z drugimi deli

- naša klasifikacijska točnost 87,65 %
- odločitvena drevesa 80,34 %
- markovske verige 81,9 %
- evalvacija preko prečnega preverjanja

Analiza napak posameznih modelov

- ročni pregled podatkov
- napake določanja mesta naglasa
  - napake privzetih besed (belgíjskem – bélgijskem, angélček – ángelček)
  - napake sestavljenih besed (hládnokŕvna – hladnokŕvna)
  - napake podobnih besed (oceníte – océnite iz poceníte)
- napake določanja tipa naglasa
  - napake večinoma pri kvaliteti naglasov (ozki in široki o ter e)

Praktična uporaba modelov

- Sloleks
  - 3 milijone besednih oblik
  - 100 000 različnih lem
  - dostopno na clarin.si

Praktična uporaba modelov

- naglaševalnik sklenjenih besedil
  - uporaba označevalnika ReLDI
  - neupoštevanje pojavitev besed

Izbrúhi na sóncu só žé vëčkrat pokazáli zóbe nášim satelítom, poslédično nášim mobílnim telefónom, navigáciji, célo eléktričnemu omréžju. Á vesóljskega vreména šë në móremo napovédati – kakó bî ga láhko, se tá téden na Blédu pogovárja okóli 70 znánstvenikov Evrópske vesóljske agéncije, ki jé sebój pripeljála svôjo näjvéčjo ikóno, británca Mátta Taylorja.

Zaključek

- strokovni prispevki
  - konvolucijske nevronske mreže so primerne za naglaševanje slovenskih besed
  - rezultati so boljši od ostalih metod strojnega učenja
  - implementirani dve praktični rešitvi
- ideje za izboljšave
  - določanje vrste naglasa po vsaki črki posebej
  - preverjanje pojavitev kombinacije mest naglasov v drugih besedah