Dve švalje

tekstove krpimo, popravljamo i prepravljamo

Cvetana Krstev i Ranka Stanković

Univerzitet u Beogradu

Seminar Društva za jezičke resurse i tehnologije – Jerteh
25. januar 2018.

Outline

Outline

Outline

Outline

Outline

Outline

Zašto tekstove treba popravljati, prepravljati i krpiti?

- popravljaju se tekstovi zbog:
  - grešaka u kucanju;
  - grešaka kod OCR;
- krpe se tekstovi zbog:
  - sistematski pogrešnog unosa teksta – "ošišana latinica";
- prepravljaju se tekstovi da bi se obavila neka "jednostavna promena":
  - promena alfabeta – ćirilica $\Leftrightarrow$ latinica;
  - promena izgovora – ekavica $\Leftrightarrow$ ijekavica.
- **2 švalje** treba da budu jedan sajt koji nudi korisnicima sve ove (i nove) usluge.

O čemu će biti reći u toku ovog izlaganja

- popravljaju se tekstovi zbog:
  - grešaka u kucanju;
  - grešaka kod OCR;
- krpe se tekstovi zbog:
  - **sistematski pogrešnog unosa teksta – "ošišana latinica" – nazovimo to vraćanje dijakritičkih znakova;**
- prepravljaju se tekstovi da bi se obavila neka "jednostavna promena":
  - promena alfabeta – ćirilica $\Leftrightarrow$ latinica;
  - promena izgovora – ekavica $\Leftrightarrow$ ijekavica.

Kako uopšte dolazi do ovog problema

- U prošlosti, zbog neadekvatne podrške latiničnim pismima koja nisu u okviru ASCII (kasnije ISO 8859-1 Latin 1);
- kod izvlačenja teksta iz PDF formata ili kod OCR (u ovim slučajevima se češće dobija tekst u kome neki dijakritički znaci nedostaju);
- kod pisanja kratkih poruka (SMS, Twitter), pa i e-poruke da bi se povećala brzina kucanja (lenjost!).

Zašto je potrebno da se problem rešava

- Da bi mogao da se dalje koristi u obradi (na primer u sistemima za glasovnu reprodukciju teksta "text-to-speech");
- Potreba (želja) da se obrađuju korpusi kratkih poruka.

Kako uopšte dolazi do ovog problema

- U prošlosti, zbog neadekvatne podrške latiničnim pismima koja nisu u okviru ASCII (kasnije ISO 8859-1 Latin 1);
- kod izvlačenja teksta iz PDF formata ili kod OCR (u ovim slučajevima se češće dobija tekst u kome neki dijakritički znaci nedostaju);
- kod pisanja kratkih poruka (SMS, Twitter), pa i e-poruke da bi se povećala brzina kucanja (lenjost!).

Zašto je potrebno da se problem rešava

- Da bi mogao da se dalje koristi u obradi (na primer u sistemima za glasovnu reprodukciju teksta “text-to-speech”);
- Potreba (želja) da se obrađuju korpusi kratkih poruka.

Kako se naziva ovaj problem u literaturi

U literaturi na engleskom jeziku se ovaj problem zove “diacritic restoration” i “diacritization”. Kada se odnosi na arapski jezik (i pismo) naziva se “vowel restoration”.

U kojim se sve jezicima javlja ovaj problem i za koje je rešavan

- francuski, hrvatski, mađarski, litvanski, rumunski, slovački, španski, turski, vijetnamski...
- arapski, hebrejski...
- grupe jezika (afrički, južnoslovenski – slovenački, hrvatski, srpski)

Kako se naziva ovaj problem u literaturi

U literaturi na engleskom jeziku se ovaj problem zove "diacritic restoration" i "diacritization". Kada se odnosi na arapski jezik (i pismo) naziva se "vowel restoration".

U kojim se sve jezicima javlja ovaj problem i za koje je rešavan

- francuski, hrvatski, mađarski, litvanski, rumunski, slovački, španski, turski, vijetnamski...
- arapski, hebrejski...
- grupe jezika (afrički, južnoslovenski – slovenački, hrvatski, srpski)

Kako se rešava ovaj problem

Malo rešenja zasnovanih na znanju i resursima

Koristio se (bar delimično) za:

- arapske glagole (El-Sadany and Hashish, 1988);
- rumunskie (Tufiș and Ceaușu, 2008);
- hrvatski (Šantić et al., 2008).

Problem idealan za statističku obradu; koriste se metode:

- koje rade na nivou karaktera i ne zahtevaju nikakve resurse:
  - litvanski (Kapociūtė-Dzikienė et al., 2017);
  - arapski (Alghamdi et al., 2010)
- koje rade na nivou reči i zahtevaju jezički model ili anotirani korpus:
  - *n*-grame – španski (Atserias et al., 2012);
  - HMM – arapski (Ibraheem, 2017), arapski/hebrejski (Gal, 2002);
  - neuronske mreže – arapski (Belinkov and Glass, 2015), vijetnamski (Pham et al., 2017).

Kako se rešava ovaj problem

Malo rešenja zasnovanih na znanju i resursima

Koristio se (bar delimično) za:

- arapske glagole (El-Sadany and Hashish, 1988);
- rumunskie (Tufiș and Ceaușu, 2008);
- hrvatski (Šantić et al., 2008).

Problem idealan za statističku obradu; koriste se metode:

- koje rade na nivou karaktera i ne zahtevaju nikakve resurse:
  - litvanski (Kapociūtė-Dzikienė et al., 2017);
  - arapski (Alghamdi et al., 2010)
- koje rade na nivou reči i zahtevaju jezički model ili anotirani korpus:
  - *n*-grame – španski (Atserias et al., 2012);
  - HMM – arapski (Ibraheem, 2017), arapski/hebrejski (Gal, 2002);
  - neuronske mreže – arapski (Belinkov and Glass, 2015), vijetnamski (Pham et al., 2017).

Rešavanje problema za srpski jezik

stranice na vebu

- Jedna stranica je `http://www.slovomajstor.com/`; nema podataka o autoru ni o korišćenoj metodi;
- Ima sigurno i drugih stranica.

U naučnoj literaturi

- Jedini rad u kome se pominje srpski jezik je rad Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, "Corpus-Based Diacritic Restoration for South Slavic Languages", sa LREC-a 2016.
- Ponuđeno je rešenje primenjeno na slovenački, hrvatski i srpski;
- U ovom rešenju se problem vraćanja dijakritika rešava kao problem prevođenja u kome se token bez dijakritika "prevodi" u token s dijakriticima;
- dobijeni fantastični rezultati (ako sam ih dobro protumačila).

Rešavanje problema za srpski jezik

stranice na vebu

- Jedna stranica je `http://www.slovomajstor.com/`; nema podataka o autoru ni o korišćenoj metodi;
- Ima sigurno i drugih stranica.

U naučnoj literaturi

- Jedini rad u kome se pominje srpski jezik je rad Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, "Corpus-Based Diacritic Restoration for South Slavic Languages", sa LREC-a 2016.
- Ponuđeno je rešenje primenjeno na slovenački, hrvatski i srpski;
- U ovom rešenju se problem vraćanja dijakritika rešava kao problem prevođenja u kome se token bez dijakritika "prevodi" u token s dijakriticima;
- dobijeni fantastični rezultati (ako sam ih dobro protumačila).

Definicija problema

Tokenizacija

U tekstu podeljenom u tokene, nas interesuju samo (jednočlane/monoleksičke) reči koje delimo u dve grupe:

- Reči koje procedura neće uzimati u obzir jer ne sadrže slova *c*, *z*, *s* (na kojima jedino mogu biti diakritički znaci), niti grupu *dj*. To su reči tipa W_a, npr. *majka*;
- Reči koje će procedura uzimati u obzir jer sadrže slova *c*, *z*, *s* ili grupu *dj*. To su reči tipa W_b;
- U reči tipa W_b spadaju reči koje sadrže kritična slova ili grupu bez obzira da li dijakritički znak ili grupa stvarno nedostaju (npr. *zvono* i *zvaka* i *podjednako* i *medjutim*);
- Za početak pretpostavljamo da u tekstu ne postoje reči sa dijakritičkim znacima (reči tipa W_c);
- I kod evaluacije će se uzimati u obzir samo W_b reči.

1. Ideja za rešavanje problema (1)

Pronaženje kandidata

Za svaku W_b reč u tekstu želimo da ponudimo listu svih kandidata – to su reči s dijakritičkim znacima koje su reči srpskog jezika:

- Ta lista može da sadrži samu polaznu reč: *liscem* $\Rightarrow$ *lišćem* (lišće), *lišcem* (lišce), *liscem* (lisac);
- ne mora da sadrži polaznu reč: *lucice* $\Rightarrow$ *lučice* (lučica), *lučiće* (lučiti);
- Ako sadrži samo polaznu reč, nema drugih kandidata, odmah se prihvata.

2. Rangiranje kandidata

Za svaku reč W_b svi njeni kandidati $(W_{b1}, W_{b2}, \ldots, W_{bn})$ treba da budu rangirani prema prema verovatnoći pojavljiavnja u tekstu. Na primer, u SrpKor *lišćem* $\rightarrow$ 265, *liscem* $\rightarrow$ 10, *lišcem* $\rightarrow$ 2.

1. Ideja za rešavanje problema (1)

Pronaženje kandidata

Za svaku W_b reč u tekstu želimo da ponudimo listu svih kandidata – to su reči s dijakritičkim znacima koje su reči srpskog jezika:

- Ta lista može da sadrži samu polaznu reč: *liscem* $\Rightarrow$ *lišćem* (lišće), *lišcem* (lišce), *liscem* (lisac);
- ne mora da sadrži polaznu reč: *lucice* $\Rightarrow$ *lučice* (lučica), *lučiće* (lučiti);
- Ako sadrži samo polaznu reč, nema drugih kandidata, odmah se prihvata.

2. Rangiranje kandidata

Za svaku reč W_b svi njeni kandidati $(W_{b1}, W_{b2}, \ldots, W_{bn})$ treba da budu rangirani prema prema verovatnoći pojavljiavnja u tekstu. Na primer, u SrpKor *lišćem* $\rightarrow$ 265, *liscem* $\rightarrow$ 10, *lišcem* $\rightarrow$ 2.

Ideja za rešavanje problema (2)

3. Izbor jednog kandidata

Za svaku reč W_b za koju postoji više kandidata W_{bi}, treba odabrati jednog i za to koristimo:

- rečnik,
- heuristiku,
- pravila (u vidu gramatika plitkog parsiranja).

4. W_b reči bez ijednog kandidata

- To će biti reči koje sistem ne prepoznaje – nisu u rečniku, vlastita imena, (regularna) derivacija, GREŠKE;
- U ovom trenutku sistem s njima ne radi ništa;
- to svakako mora da se promeni.

Ideja za rešavanje problema (2)

3. Izbor jednog kandidata

Za svaku reč W_b za koju postoji više kandidata W_{bi}, treba odabrati jednog i za to koristimo:

- rečnik,
- heuristiku,
- pravila (u vidu gramatika plitkog parsiranja).

4. W_b reči bez ijednog kandidata

- To će biti reči koje sistem ne prepoznaje – nisu u rečniku, vlastita imena, (regularna) derivacija, GREŠKE;
- U ovom trenutku sistem s njima ne radi ništa;
- to svakako mora da se promeni.

Rečnik za potrebe vraćanja dijakritičkih znakova (1)

Transformacija Srpskog morfološkog rečnika (SMD) DELAF i DELACF

- Iz SMD se vade svi oblici reči koji su tipa W_c (sa dijakritičkim znaicma) i W_b (sa potencijalno nedostajućim dijakritičkim znacima);
- Sve reči tipa W_c se prevode u reči tipa W_b ukidanjem dijakritičkih znakova;
- Sve nepotrebne informacije se uklanjaju (lema, vrsta reči, itd...);
- Svi tako dobijeni isti oblici se spajaju.

Primer

liscem,lisac.N+Zool:ms6v liscem,.X+CR=liscem
lišćem,lišće.N+Conc:ns6q liscem,.X+CR=lišćem ⇒ **SMD_DR rečnik**
lišcem,lišce.N+Dem:ns6q liscem,.X+CR=lišcem

liscem,.X+CR=liscem_lišćem_liscem

Rečnik za potrebe vraćanja dijakritičkih znakova (1)

Transformacija Srpskog morfološkog rečnika (SMD) DELAF i DELACF

- Iz SMD se vade svi oblici reči koji su tipa W_c (sa dijakritičkim znaicma) i W_b (sa potencijalno nedostajućim dijakritičkim znacima);
- Sve reči tipa W_c se prevode u reči tipa W_b ukidanjem dijakritičkih znakova;
- Sve nepotrebne informacije se uklanjaju (lema, vrsta reči, itd...);
- Svi tako dobijeni isti oblici se spajaju.

Primer

liscem,lisac.N+Zool:ms6v liscem,.X+CR=liscem
lišćem,lišće.N+Conc:ns6q liscem,.X+CR=lišćem ⇒ **SMD_DR rečnik**
lišcem,lišce.N+Dem:ns6q liscem,.X+CR=lišcem

liscem,.X+CR=liscem_lišćem_liscem

Rečnik za potrebe vraćanja dijakritičkih znakova (2)

Postupanje s malim i velikim slovima

- Velika slova se slažu samo s velikim slovima;
- Mala slova se slažu i s malim i s velikim slovima.

Primer

liže,lizati.V+Imperf:Psz lize,.X+CR=liže
Lize,Liza.N+NProp:fs2v Lize,.X+CR=Lize Lize,Liza.X+CORR=Lize_liže

Rečnik za potrebe vraćanja dijakritičkih znakova (2)

Postupanje s malim i velikim slovima

- Velika slova se slažu samo s velikim slovima;
- Mala slova se slažu i s malim i s velikim slovima.

Primer

liže,lizati.V+Imperf:Psz lize,.X+CR=liže
Lize,Liza.N+NProp:fs2v Lize,.X+CR=Lize Lize,Liza.X+CORR=Lize_liže

Rečnik za potrebe vraćanja dijakritičkih znakova (3)

Dimenzije rečnika

- Rečnik ima 943.804 unosa. Od toga:
  - sa jednom ponudom 897.077 (95,05%);
  - sa dve ponude 41.444(4,38%);
  - sa tri ponude 3.585 (0,38%);
  - sa više od tri ponude 569 (0,06%).

Maksimalan broj kandidata je 8

To je sledeći unos iz rečnika:
Celice,.N+CR=Čeliče_Celiće_Ćeliće_Čeliće_
čeliče_ćelice_celice_celiće
Oblici prezimena *Čelik*, *Ćelić*, *Ćelić*, *Čelić*, imenica *čelik*, *celica*, *ćelica* i glagola *čeličiti*, *celiti*.

Rečnik za potrebe vraćanja dijakritičkih znakova (3)

Dimenzije rečnika

- Rečnik ima 943.804 unosa. Od toga:
  - sa jednom ponudom 897.077 (95,05%);
  - sa dve ponude 41.444(4,38%);
  - sa tri ponude 3.585 (0,38%);
  - sa više od tri ponude 569 (0,06%).

Maksimalan broj kandidata je 8

To je sledeći unos iz rečnika:
Celice,.N+CR=Čeliče_Celiće_Ćeliće_Čeliće_
čeliče_ćelice_celice_celiće
Oblici prezimena *Čelik*, *Celić*, *Ćelić*, *Čelić*, imenica *čelik*, *celica*, *ćelica* i glagola *čeličiti*, *celiti*.

Rečnik za rangiranje kandidata

Informacije o frekvencijama

- Ove informacije su izračunate na osnovu dela SrpKor od oko 108 miliona reči;
- određen je ukupan broj tokena-reči ($totalNumTokens$);
- za tokene zapisane velikim slovom računate su samo frekvencije tokena koji su zapisani velikim slovom;
- za tokene zapisane malim slovom računata je ukupna frekvencija.

Relativne frekvencije

$$relFreq = Round\left(\frac{freq \cdot 10000000}{totalNumTokens + 0.5}, 0\right)$$

Relativna frekvencija za 0 je 0, za 1–10 je 1, 11–21 je 2, 22–32 je 3,...
Maksimalna apsolutna frekvencija je 3,706.356, $relFreq = 340596$.
Za polileksičke jedinice frekvencija nije računata.

Rečnik za rangiranje kandidata

Informacije o frekvencijama

- Ove informacije su izračunate na osnovu dela SrpKor od oko 108 miliona reči;
- određen je ukupan broj tokena-reči ($totalNumTokens$);
- za tokene zapisane velikim slovom računate su samo frekvencije tokena koji su zapisani velikim slovom;
- za tokene zapisane malim slovom računata je ukupna frekvencija.

Relativne frekvencije

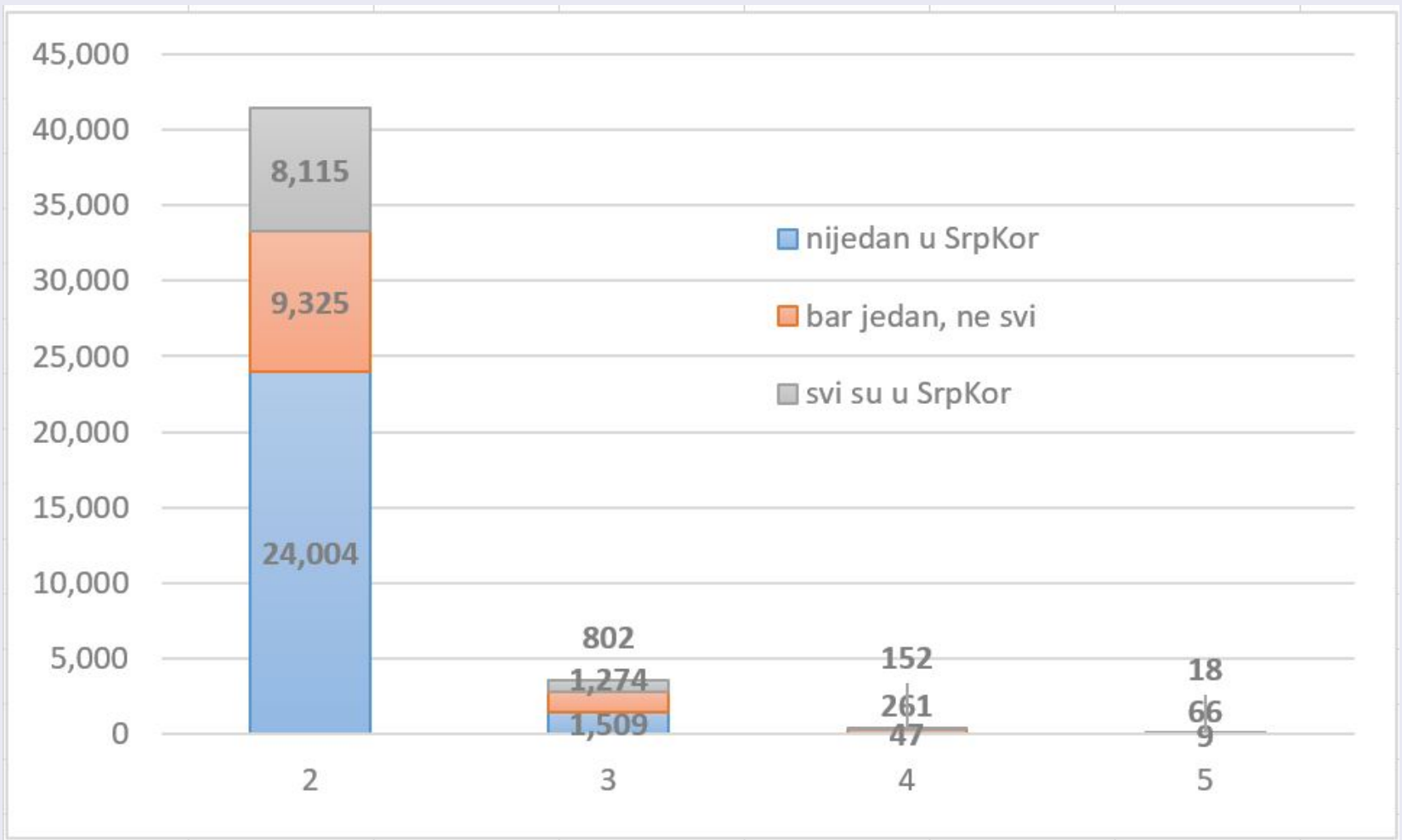
$$relFreq = Round\left(\frac{freq \cdot 10000000}{totalNumTokens + 0.5}, 0\right)$$

Relativna frekvencija za 0 je 0, za 1–10 je 1, 11–21 je 2, 22–32 je 3,...
Maksimalna apsolutna frekvencija je 3,706.356, $relFreq = 340596$.
Za polileksičke jedinice frekvencija nije računata.

Distribucija unosa u SMD_DR prema broju kandidata i pojavljivanju u SrpKor

Pojavljivanje u SrpKor	**Broj Kandidata**					
	1	**2**	**3**	**4**	**5**	>6
nijedan u SrpKor	742.941	24.004	1.509	47	9	2
%	82,82	57,92	42,09	10,22	9,68	12,50
bar jedan, ne svi		9.325	1.274	261	66	12
%		22,50	35,54	56,74	70,97	75,00
svi su u SrpKor	154.136	8.115	802	152	18	2
%	17,18	19,58	22,37	33,04	19,35	12,5
Total	897.077	41.444	3.585	460	93	16

Distribucija broja kandidata (kad ih ima više od 1)



Koliko je stepen dvosmislenosti reči bez dijakritičkih znakova

Najteži slučajevi

- 147 unosa sa dva kandidata od kojih svi imaju $relFreq \geq 100$;
  - 12 unosa sa dva kandidata od kojih svi imaju $relFreq \geq 1000$;
- 4 unosa sa tri kandidata od kojih svi imaju $relFreq \geq 100$.

Najteži slučajevi

- reci,.X+CR=reci(237)_reči(2607)_reći(1448)
- sto,.X+CR=što(36850)_sto(1268)
- nas,.X+CR=nas(5623)_naš(3528)

Koliko je stepen dvosmislenosti reči bez dijakritičkih znakova

Najteži slučajevi

- 147 unosa sa dva kandidata od kojih svi imaju $relFreq \geq 100$;
  - 12 unosa sa dva kandidata od kojih svi imaju $relFreq \geq 1000$;
- 4 unosa sa tri kandidata od kojih svi imaju $relFreq \geq 100$.

Najteži slučajevi

- reci,.X+CR=reci(237)_reči(2607)_reći(1448)
- sto,.X+CR=što(36850)_sto(1268)
- nas,.X+CR=nas(5623)_naš(3528)

Osnovne ideje

Obrada teksta

- Obrada teksta podrazumeva:
  - tokenizaciju
  - pridruživanje skupova rečničkih informacija svakom obliku reči (lema, vrsta reči, gramatičke (i druge) informacije;
- Za ovaj posao se koristi **Unitex**, paket za obradu korpusa podržanu leksičkim resursima (`http://unitexgramlab.org/`) i SMD.
- Skup rečničkih informacija koji se pridružuje svakom obliku reči može da bude:
  - prazan – oblik reči nije u rečniku (npr. *lize*);
  - može da ima jednu ili više informacija, npr.:
    - **reci,reka.N:fs7q:fs3q**
    - **reci,redak.N:mp5q:mp1q**
    - **reci,reći.V:Yys**
  - (a ipak da ne sadrži pravu, koja bi mogla biti **reči,reč.N:fp5q:fp4q**...)

Pridruživanje kandidata

Morfološki režim – Rečnici SMD_DR se **NE** koriste za obradu teksta!

- Za ovaj posao se takođe koristi **Unitex** i SMD_DR.
- Koristi se morfološki režima rada Unitex i SMD_DR kao morfološki rečnici;
- Ovaj režim dozvoljava da se za reč prepoznatu u morfološkom rečniku izvade željene informacije iz rečnika.

Kako to radi

- Neka je u tekstu token *reci* koji pripada tipu W_b, dakle nad njim će naša procedura raditi;
- On će dobiti iz SMD razne rečničke interpretacije (prethodni slajd), ali to u ovom trenutku nije bitno;
- Pošto je *reci* tipa W_b, reč se traži i pronalazi u rečniku SMD_DR i iz njega se preuzima vrednost atributa **+CR**, u ovom slučaju **reci(237)_reči(2607)_reći(1448)**.

Pridruživanje kandidata

Morfološki režim – Rečnici SMD_DR se **NE** koriste za obradu teksta!

- Za ovaj posao se takođe koristi **Unitex** i SMD_DR.
- Koristi se morfološki režima rada Unitex i SMD_DR kao morfološki rečnici;
- Ovaj režim dozvoljava da se za reč prepoznatu u morfološkom rečniku izvade željene informacije iz rečnika.

Kako to radi

- Neka je u tekstu token *reci* koji pripada tipu W_b, dakle nad njim će naša procedura raditi;
- On će dobiti iz SMD razne rečničke interpretacije (prethodni slajd), ali to u ovom trenutku nije bitno;
- Pošto je *reci* tipa W_b, reč se traži i pronalazi u rečniku SMD_DR i iz njega se preuzima vrednost atributa **+CR**, u ovom slučaju **reci(237)_reči(2607)_reći(1448)**.

Šta su kaskade

Kaskade u Unitex-u

- Kaskade su sekvence transduktora koji se primenjuju na tekst u utvrđenom redosledu;
- Svakim prolaskom tekst se transformiše – izlaz iz jednog transduktora je ulaz u sledeći;
- Svaki transduktor u tekstu može nešto da doda (*merge*) ili promeni (*replace*), a transduktori se mogu i iterativno propuštati dok se ne dođe do tačke kada više nema promene.
- Poenta kaskada je da jedan transduktor može deo teksta da zaokruži (npr. rešen problem) tako da ga sledeći transduktor više ne dira, ali može da ga koristi.
- U Unitex-u deo koji radi s kaskadama se zove CaSys (Natalie Fribgurger & Denis Maurel).

Dve kaskade

Prva kaskada

- Polazi od teksta koji je obrađen (primenjeni su rečnici);
- Izvlači informacije o kandidatima za korekciju W_b reči;
- Rešava neke jasne slučajeve (trebalo bi da budu slučajevi koji neće uneti pogrešnu zamenu niti neku propustiti).

Druga kaskada

- Postepeno se prihvataju neki kandidati, a odbacuju drugi s ciljem da svaka W_b reč dobije jednog kandidata kojim se onda zamenjuje;
- Neki koraci ove kaskade se mogu izostaviti ili zameniti što zavisi od konkretnog slučaja ili korisnika.

Dve kaskade

Prva kaskada

- Polazi od teksta koji je obrađen (primenjeni su rečnici);
- Izvlači informacije o kandidatima za korekciju W_b reči;
- Rešava neke jasne slučajeve (trebalo bi da budu slučajevi koji neće uneti pogrešnu zamenu niti neku propustiti).

Druga kaskada

- Postepeno se prihvataju neki kandidati, a odbacuju drugi s ciljem da svaka W_b reč dobije jednog kandidata kojim se onda zamenjuje;
- Neki koraci ove kaskade se mogu izostaviti ili zameniti što zavisi od konkretnog slučaja ili korisnika.

Prva kaskada – korak po korak (1)

Korak 1 – Veliko slovo

U ovom koraku se beleže sve pozicije u tekstu gde imamo veliko slovo, jer zamenom iz rečnika može na tom mestu da se dobije malo slovo (jer nije vlastito ime), pa to na kraju treba vratiti u veliko slovo.

Korak 2 – *ad hoc* pravila razrešavanja

- Ova pravila razrešavaju neke slučajeve na osnovu konteksta (u tom kontekstu se gledaju reči koje mogu da imaju interpretaciju!):
- Primer pravila koja potvrđuju **sto**:
  - **sto** iza čega sledi **puta** ili **posto**;
  - **sto** iza čega sledi broj (kao u **sto hiljada**);
  - **sto** iza čega sledi fraza pridev/imenica u genitivu plurala;

Prva kaskada – korak po korak (1)

Korak 1 – Veliko slovo

U ovom koraku se beleže sve pozicije u tekstu gde imamo veliko slovo, jer zamenom iz rečnika može na tom mestu da se dobije malo slovo (jer nije vlastito ime), pa to na kraju treba vratiti u veliko slovo.

Korak 2 – *ad hoc* pravila razrešavanja

- Ova pravila razrešavaju neke slučajeve na osnovu konteksta (u tom kontekstu se gledaju reči koje mogu da imaju interpretaciju!):
- Primer pravila koja potvrđuju **sto**:
  - **sto** iza čega sledi **puta** ili **posto**;
  - **sto** iza čega sledi broj (kao u **sto hiljada**);
  - **sto** iza čega sledi fraza pridev/imenica u genitivu plurala;

Prva kaskada – korak po korak (2)

Korak 3 – trigrami

- Koristi se 30 najfrekventnijih trigrama (prema SrpKor) u okviru kojih je bar jedna W_b reč. Na primer:
  - **sto se tice** $\Rightarrow$ **što se tiče**;
  - **na taj nacin** $\Rightarrow$ **na taj način**;

Korak 4 – bigrami

- Koristi se 50 najfrekventnijih bigrama (prema SrpKor) u okviru kojih je bar jedna W_b reč. Na primer:
  - **sto ce** $\Rightarrow$ **što će**;
  - **je takodje** $\Rightarrow$ **je takođe**;

Prva kaskada – korak po korak (2)

Korak 3 – trigrami

- Koristi se 30 najfrekventnijih trigrama (prema SrpKor) u okviru kojih je bar jedna W_b reč. Na primer:
  - **sto se tice** $\Rightarrow$ **što se tiče**;
  - **na taj nacin** $\Rightarrow$ **na taj način**;

Korak 4 – bigrami

- Koristi se 50 najfrekventnijih bigrama (prema SrpKor) u okviru kojih je bar jedna W_b reč. Na primer:
  - **sto ce** $\Rightarrow$ **što će**;
  - **je takodje** $\Rightarrow$ **je također**;

Prva kaskada – korak po korak (3)

Korak 5 – Polileksičke jedinice

- U ovom koraku se kao morfološki rečnik koristi rečnik SMD_DR polileksičkih jedinica.
- Pretpostavka je da su polileksičke jedinice uglavnom nedvosmislene. Na primer,
  - **kljucne reci** $\Rightarrow$ **ključne reči**;
  - **kozne obuce** $\Rightarrow$ **kožne obuće**;

Korak 6 – Monoleksičke jedinice

- U ovom koraku se kao morfološki rečnik koristi rečnik SMD_DR monoleksičkih jedinica.
- Zamenjuje se W_b reč (ako prethodnim koracima nije rešena) bilo da:
  a) ta reč nije u rečniku SMD_DR (primer **laze**);
  b) ta reč jeste u rečniku SMD_DR (primer **reci**).

Prva kaskada – korak po korak (3)

Korak 5 – Polileksičke jedinice

- U ovom koraku se kao morfološki rečnik koristi rečnik SMD_DR polileksičkih jedinica.
- Pretpostavka je da su polileksičke jedinice uglavnom nedvosmislene. Na primer,
  - **kljucne reci** $\Rightarrow$ **ključne reči**;
  - **kozne obuce** $\Rightarrow$ **kožne obuće**;

Korak 6 – Monoleksičke jedinice

- U ovom koraku se kao morfološki rečnik koristi rečnik SMD_DR monoleksičkih jedinica.
- Zamenjuje se W_b reč (ako prethodnim koracima nije rešena) bilo da:
  a) ta reč nije u rečniku SMD_DR (primer **laze**);
  b) ta reč jeste u rečniku SMD_DR (primer **reci**).

Primer rada prve kaskade

Sve što uradi prva kaskada ostaje zabeleženo u tekstu. Odabirom koraka iz druge kaskade korisnik će odlučiti šta ostaje u tekstu, šta se dalje transformiše, a šta se uklanja.

polazni tekst	obeležen tekst
Jer je imao dovoljno vremena da spreci zlocin cak i nakon reci kojima je podstrekivao sina.	1_Jer je imao dovoljno vremena da 6a_(spreči(302)_spreći(0)) 6a_(zločin(456)) 4_(čak i) 2_(nakon reči) kojima je podstrekivao 6b_(sina(518)_šina(54)).
U novinama vise nije bilo ni reci o ratnoj steti.	1_U novinama 6b_(vise(35)_više(17628)) 2_(nije bilo ni reči o) 5_(ratnoj šteti(0)).

Primer rada prve kaskade

Sve što uradi prva kaskada ostaje zabeleženo u tekstu. Odabirom koraka iz druge kaskade korisnik će odlučiti šta ostaje u tekstu, šta se dalje transformiše, a šta se uklanja.

polazni tekst	**obeležen tekst**
Jer je imao dovoljno vremena da spreci zlocin cak i nakon reci kojima je podstrekivao sina.	1_Jer je imao dovoljno vremena da 6a_(spreči(302)_spreći(0)) 6a_(zločin(456)) 4_(čak i) 2_(nakon reči) kojima je podstrekivao 6b_(sina(518)_šina(54)).
U novinama vise nije bilo ni reci o ratnoj steti.	1_U novinama 6b_(vise(35)_više(17628)) 2_(nije bilo ni reči o) 5_(ratnoj šteti(0)).

Druga kaskada - korak po korak (1)

Ponovna obrada teksta rečnicima SMD

Obeleženom tekstu se ponovo pridružuju informacije iz SMD. To je potrebno jer će neka pravila odlučivanja to zahtevati.

Korak 1 – prvo čišćenje

Uklanjaju se neke oznake, koje se tiču oznake velikih slova ako je reč ostala zapisana velikim slovom: **1_Mocart ⇒ Mocart**;

Korak 2 – uklanjaju se kandidati koji nisu u SrpKor

U ovom trenutku samo ako postoji i kandidat koji jeste u SrpKor:

- **6b_(jezikom(216)_ježikom(0)) ⇒ 6b_(jezikom(216))**;
- **6b_(srcu(235)_srču(2)_srću(0)) ⇒ 6b_(srcu(235)_srču(2))**;
- **6b_(lišče(0)_lisce(1)_lišce(1)_lišće(60)_lišče(0)) ⇒ 6b_(lisce(1)_lišce(1)_lišće(60))**.

Druga kaskada - korak po korak (1)

Ponovna obrada teksta rečnicima SMD

Obeleženom tekstu se ponovo pridružuju informacije iz SMD. To je potrebno jer će neka pravila odlučivanja to zahtevati.

Korak 1 – prvo čišćenje

Uklanjaju se neke oznake, koje se tiču oznake velikih slova ako je reč ostala zapisana velikim slovom: **1_Mocart** $\Rightarrow$ **Mocart**;

Korak 2 – uklanjaju se kandidati koji nisu u SrpKor

U ovom trenutku samo ako postoji i kandidat koji jeste u SrpKor:

- **6b_(jezikom(216)_ježikom(0))** $\Rightarrow$ **6b_(jezikom(216))**;
- **6b_(srcu(235)_srču(2)_srću(0))** $\Rightarrow$ **6b_(srcu(235)_srču(2))**;
- **6b_(lišče(0)_lisce(1)_lišce(1)_lišće(60)_lišče(0))** $\Rightarrow$ **6b_(lisce(1)_lišce(1)_lišće(60))**.

Druga kaskada - korak po korak (1)

Ponovna obrada teksta rečnicima SMD

Obeleženom tekstu se ponovo pridružuju informacije iz SMD. To je potrebno jer će neka pravila odlučivanja to zahtevati.

Korak 1 – prvo čišćenje

Uklanjaju se neke oznake, koje se tiču oznake velikih slova ako je reč ostala zapisana velikim slovom: **1_Mocart** $\Rightarrow$ **Mocart**;

Korak 2 – uklanjaju se kandidati koji nisu u SrpKor

U ovom trenutku samo ako postoji i kandidat koji jeste u SrpKor:

- **6b_(jezikom(216)_ježikom(0))** $\Rightarrow$ **6b_(jezikom(216))**;
- **6b_(srcu(235)_srčcu(2)_srćcu(0))** $\Rightarrow$ **6b_(srcu(235)_srčcu(2))**;
- **6b_(lisčce(0)_lisce(1)_lišce(1)_lišće(60)_lišče(0))** $\Rightarrow$ **6b_(lisce(1)_lišce(1)_lišće(60))**.

Druga kaskada – korak po korak (2)

Korak 3 – prihvatanje jedinstvenih kandidata

Bilo da su potvrđeni u SrpKor ili ne:

- **6a_(žalosti(92)) ⇒ žalosti**;
- **6a_(djevojaštva(0)) ⇒ djevojaštva**;

Korak 4 – biranje kandidata koji se mnogo češće pojavljuje

Ova kaskada može da se menja u skladu s tim šta znači "mnogo češće". Na primer, ako je to reda veličine 100 puta:

- **6b_(bas(65)_baš(2444)) ⇒ 6b_(baš(2444))**;
- **6a_(času(91)_času(272)_ćasu(2)) ⇒ 6a_(času(91)_času(272))**;
- **6b_(lice(1821)_liče(79)_liće(1)) ⇒ 6b_(lice(1821)_liće(1)) ⇒ 6b_(lice(1821))**.

Druga kaskada – korak po korak (2)

Korak 3 – prihvatanje jedinstvenih kandidata

Bilo da su potvrđeni u SrpKor ili ne:

- **6a_(žalosti(92))** ⇒ **žalosti**;
- **6a_(djevojaštva(0))** ⇒ **djevojaštva**;

Korak 4 – biranje kandidata koji se mnogo češće pojavljuje

Ova kaskada može da se menja u skladu s tim šta znači "mnogo češće". Na primer, ako je to reda veličine 100 puta:

- **6b_(bas(65)_baš(2444))** ⇒ **6b_(baš(2444))**;
- **6a_(času(91)_času(272)_ćasu(2))** ⇒ **6a_(času(91)_času(272))**;
- **6b_(lice(1821)_liče(79)_liće(1))** ⇒ **6b_(lice(1821)_liće(1))** ⇒ **6b_(lice(1821))**.

Druga kaskada – korak po korak (3)

Korak 5 – prva runda biranja između više kandidata

Gleda se u konekst reči s više kandidata. Ako su u tom kontekstu W_a reči ili potvrđene reči W_b ili W_c one mogu da odluče:

- pridev je potvrđen, a iza njega slede kandidati među kojima je i imenica sa kojima se pridev slaže u rodu, broju i padežu, onda se imenica potvrđuje:
  - **slatkih 6b_(rijeci(50)_riječi(44)) ⇒ slatkih 6b_(riječi(44))**;
- Ako je imenica potvrđena, a ispred nje su kandidati među kojima je i pridev koji se s imenicom slaže u rodu, broju i padežu, onda se pridev potvrđuje:
  - **6b_(čelo(212)_celo(181)) popodne ⇒ 6b_(celo(181)) popodne**;
- Ako je predlog potvrđen, a iz njega slede kandidati među kojoma su pridev, imenica ili zamenica koja je u padežu koji predlog zahteva, oni se potvrđuju:
  - **ni u 6b_(muci(25)_muči(96)) ⇒ ni u 6b_(muci(25))**.
- Tamo gde ima potvrđenih kandidata oni ostaju, ostali se brišu.

Druga kaskada – korak po korak (4)

Korak 6 – ponavlja se prihvatanje jedinstvenih

Jer su neki postali jedinstveni u prethodnim koracima.

Korak 7 - druga runda biranja između više kandidata

Opet se gleda u kontekst:

- rešavaju se neki specijalni slučajevi (**što/sto**, **naš/nas**, itd):
  - **6b_(što(36850)_sto(1268)) ne činimo ⇒ što ne činimo** ;
- Rečca **se** iza koje sledi povratni glagol:
  - **da se 6a_(suši(30)_šuši(1)); ⇒ da se suši**;
- Pomoćni glagol **će/neće** iz koga sledi glagol u infinitivu:
  - **da se neće 6b_(obuci(49)_obući(30)_obući(5)) ⇒ da se neće obući**;
- Rečca **ne** iza koje sledi lični glagolski oblik;
  - **ne 6b_(tući(67)_tući(12)_tuci(1)) me ⇒ ne tuci me**.

Druga kaskada – korak po korak (4)

Korak 6 – ponavlja se prihvatanje jedinstvenih

Jer su neki postali jedinstveni u prethodnim koracima.

Korak 7 - druga runda biranja između više kandidata

Opet se gleda u kontekst:

- rešavaju se neki specijalni slučajevi (**što/sto**, **naš/nas**, itd):
  - **6b_(što(36850)_sto(1268)) ne činimo ⇒ što ne činimo** ;
- Rečca **se** iza koje sledi povratni glagol:
  - **da se 6a_(suši(30)_šuši(1)); ⇒ da se suši**;
- Pomoćni glagol **će/neće** iz koga sledi glagol u infinitivu:
  - **da se neće 6b_(obuci(49)_obući(30)_obući(5)) ⇒ da se neće obući**;
- Rečca **ne** iza koje sledi lični glagolski oblik;
  - **ne 6b_(tući(67)_tući(12)_tuci(1)) me ⇒ ne tuci me**.

Druga kaskada – korak po korak (5)

Korak 8 - poslednji korak

- Vraćanje svih velikih slova;
- Izbacivanje duplikata (malo/veliko slovo);
- Korak koji može a ne mora da se uradi: ako i dalje ima višestrukih kandidata, onda:
  - Ako je u listi kandidata i polazna reč, bez dijakritičkih znakova, ona se bira;
  - Ako u listi kandidata svi imaju dijakritičke znake, bira se onaj koji ima veću frekvenciju (pa koliko bilo).

Kako doprinose koraci obe kaskade

Prva kaskada			Druga kaskada		
Korak	**Ek** 2.024	**Ijk** 1.930	**Korak**	**Ek** 2.024	**Ijk** 1.930
1 (v. slovo)	185	163	1 (čišćenje)	164	136
2 (više kandidata)	8	12	2 ($W_c \notin \mathrm{SrpKor}$)	23	20
3 (trigrami)	1	1	3 (W_c jedinstven)	210	311
4 (bigrami)	9	16	4 (frequency)	55	64
5 (MWU)	3	0	5 (više kandidata 1)	12	5
6 ($W_b \notin \mathrm{SMD}$)	201	257	6 (W_c jedinstven)	96	78
7 ($W_b \in \mathrm{SMD}$)	138	145	7 (više kandidata 2)	2	4
			8 (završno čišćenje)	64	91

- Brojevi daju izmene po koraku kaskade, ali svaka izmena može da se odnosi na više tokena, i na svaki token, može da utiče više koraka;
- Doprinos rečnika, čak i u ovako kratkim tekstovima bilo je jedinstvenih kandidata sa $relFreq = 0$: 5 u Ek tekstu, 7 u Ijk tekstu.

Kako doprinose koraci obe kaskade

Prva kaskada			Druga kaskada		
Korak	**Ek** 2.024	**Ijk** 1.930	**Korak**	**Ek** 2.024	**Ijk** 1.930
1 (v. slovo)	185	163	1 (čišćenje)	164	136
2 (više kandidata)	8	12	2 ($W_c \notin \mathrm{SrpKor}$)	23	20
3 (trigrami)	1	1	3 (W_c jedinstven)	210	311
4 (bigrami)	9	16	4 (frequency)	55	64
5 (MWU)	3	0	5 (više kandidata 1)	12	5
6 ($W_b \notin \mathrm{SMD}$)	201	257	6 (W_c jedinstven)	96	78
7 ($W_b \in \mathrm{SMD}$)	138	145	7 (više kandidata 2)	2	4
			8 (završno čišćenje)	64	91

- Brojevi daju izmene po koraku kaskade, ali svaka izmena može da se odnosi na više tokena, i na svaki token, može da utiče više koraka;
- Doprinos rečnika, čak i u ovako kratkim tekstovima bilo je jedinstvenih kandidata sa $relFreq = 0$: 5 u Ek tekstu, 7 u Ijk tekstu.

Prva verzija evaluacije

Tekst za evaluaciju

- Korišćen je deo korpusa SrpKor, *Politikin* kulturni dodatak iz 2001. godine;
  - Možda je to bila greška, trebalo je uzeti novi tekst!;
- Tekst je imao 557.679 oblika reči (tokena), 84.041 različitih (tipova);
- Tekst je imao 283.951 reči tipa W_b (na koje procedura "deluje"), 57,927 različitih.

Postupak evaluacije

- U tekstu su automatski skinuti svi dijakritički znakovi;
- Primenjena je procedura za vraćanje dijakritičkih znakova;
- Napravljene su vreće reči oba teksta (Bag-Of-Words, BOW);
- Te dve vreće smo poredili (samo reči W_b iz originalnog teksta).

Prva verzija evaluacije

Tekst za evaluaciju

- Korišćen je deo korpusa SrpKor, *Politikin* kulturni dodatak iz 2001. godine;
  - Možda je to bila greška, trebalo je uzeti novi tekst!;
- Tekst je imao 557.679 oblika reči (tokena), 84.041 različitih (tipova);
- Tekst je imao 283.951 reči tipa W_b (na koje procedura "deluje"), 57,927 različitih.

Postupak evaluacije

- U tekstu su automatski skinuti svi dijakritički znakovi;
- Primenjena je procedura za vraćanje dijakritičkih znakova;
- Napravljene su vreće reči oba teksta (Bag-Of-Words, BOW);
- Te dve vreće smo poredili (samo reči W_b iz originalnog teksta).

Rezultati prve evaluacije (1)

	TP	TN	FP	FN	Total
tokeni	21,112	33,805	423	2,587	57,927
relativno	0.364	0.584	0.007	0.045	1
frekvencije	93,703	172,727	3,536	13,985	283,951
relativno	0.330	0.608	0.012	0.049	1

	P	*R*	*Acc*	*F1*
tokeni	0.980	0.891	0.948	0.933
frekvencije	0.964	0.870	0.938	0.915

$P = tp/(tp + fp)$ $\quad R = tp/(tp + fn)$

$Acc = (tp + tn)/(tp + tn + fp + fp)$ $\quad F_1 = 2 \cdot P \cdot R/(P + R)$

Rezultati prve evaluacije (1)

	TP	TN	FP	FN	Total
tokeni	21,112	33,805	423	2,587	57,927
relativno	0.364	0.584	0.007	0.045	1
frekvencije	93,703	172,727	3,536	13,985	283,951
relativno	0.330	0.608	0.012	0.049	1

	P	*R*	*Acc*	*F1*
tokeni	0.980	0.891	0.948	0.933
frekvencije	0.964	0.870	0.938	0.915

$$P = tp/(tp + fp) \qquad R = tp/(tp + fn)$$
$$Acc = (tp + tn)/(tp + tn + fp + fp) \qquad F_1 = 2 \cdot P \cdot R/(P + R)$$

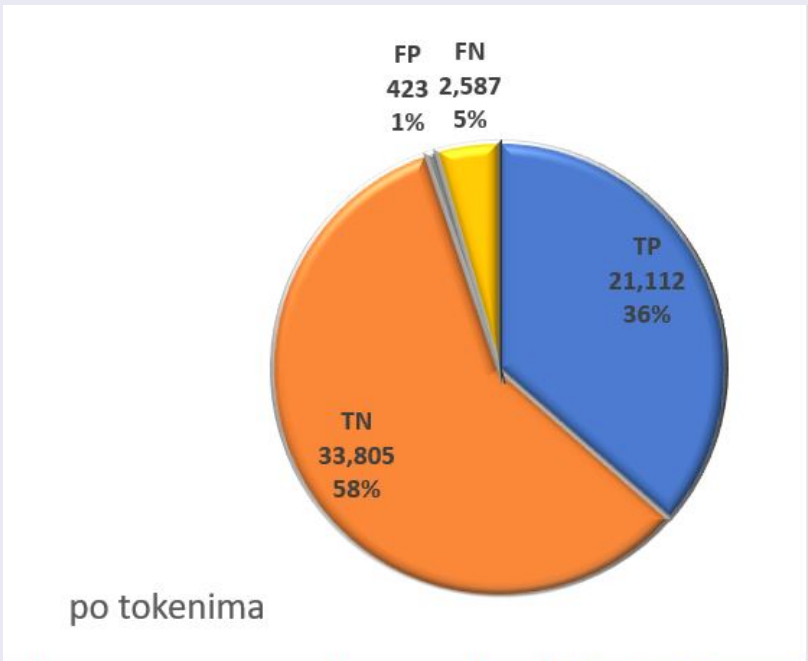
Rezultati prve evaluacije (1)

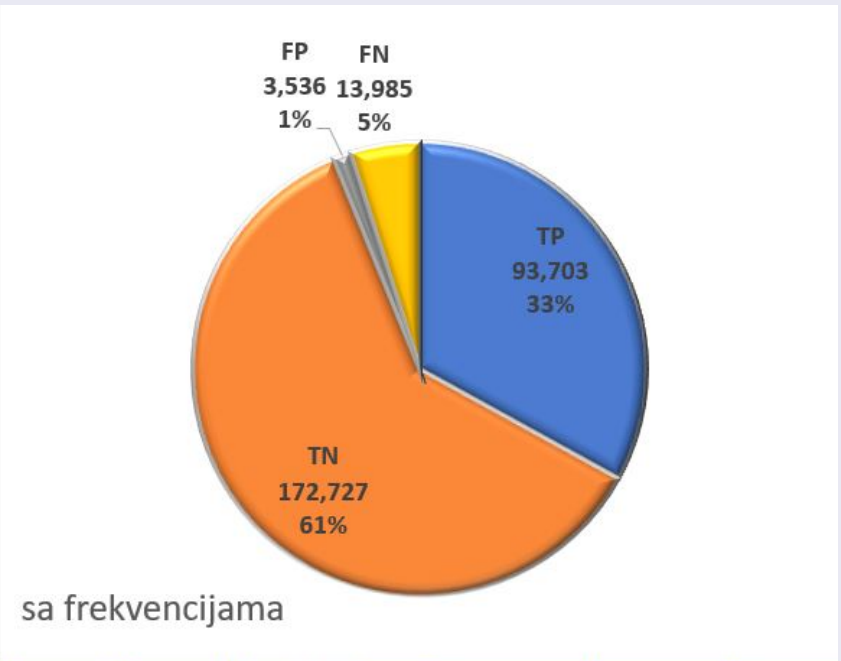
	TP	TN	FP	FN	Total
tokeni	21,112	33,805	423	2,587	57,927
relativno	0.364	0.584	0.007	0.045	1
frekvencije	93,703	172,727	3,536	13,985	283,951
relativno	0.330	0.608	0.012	0.049	1

	P	*R*	*Acc*	*F1*
tokeni	0.980	0.891	0.948	0.933
frekvencije	0.964	0.870	0.938	0.915

$$P = tp/(tp + fp) \qquad R = tp/(tp + fn)$$
$$Acc = (tp + tn)/(tp + tn + fp + fp) \qquad F_1 = 2 \cdot P \cdot R/(P + R)$$

Rezultati prve evaluacije (2)





Druga evaluacija

Postupak

- Korišćeno je 65 različitih tekstova, različite dužine;
- Kao i ranije, automatski su skinuti dijakritički znakovi;
- Poravnate su rečenice "ošišanog" i restauriranog teksta, pa smo poredili reč po reč.

	P	R	Acc	F_1
tokeni – prosek	0.986	0.939	0.969	0.962
najveći	0.997	0.961	0.981	0.977
najmanji	0.916	0.895	0.944	0.930
frekvencije – prosek	0.989	0.949	0.978	0.968
najveći	1.000	0.984	0.994	0.992
prosek	0.958	0.885	0.950	0.929

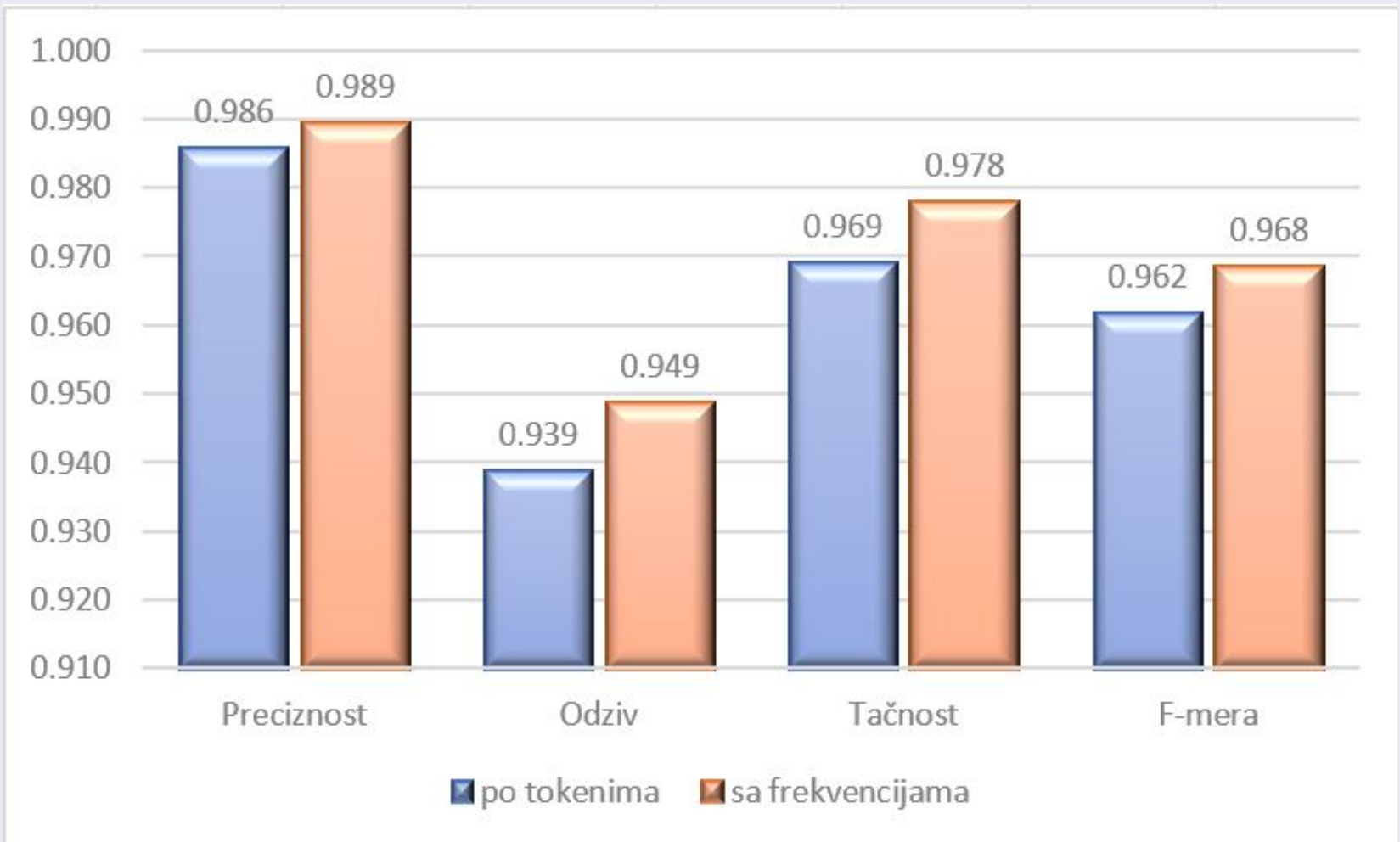
Druga evaluacija

Postupak

- Korišćeno je 65 različitih tekstova, različite dužine;
- Kao i ranije, automatski su skinuti dijakritički znakovi;
- Poravnate su rečenice "ošišanog" i restauriranog teksta, pa smo poredili reč po reč.

	P	R	Acc	F_1
tokeni – prosek	0.986	0.939	0.969	0.962
najveći	0.997	0.961	0.981	0.977
najmanji	0.916	0.895	0.944	0.930
frekvencije – prosek	0.989	0.949	0.978	0.968
najveći	1.000	0.984	0.994	0.992
prosek	0.958	0.885	0.950	0.929

Rezultati druge evaluacije



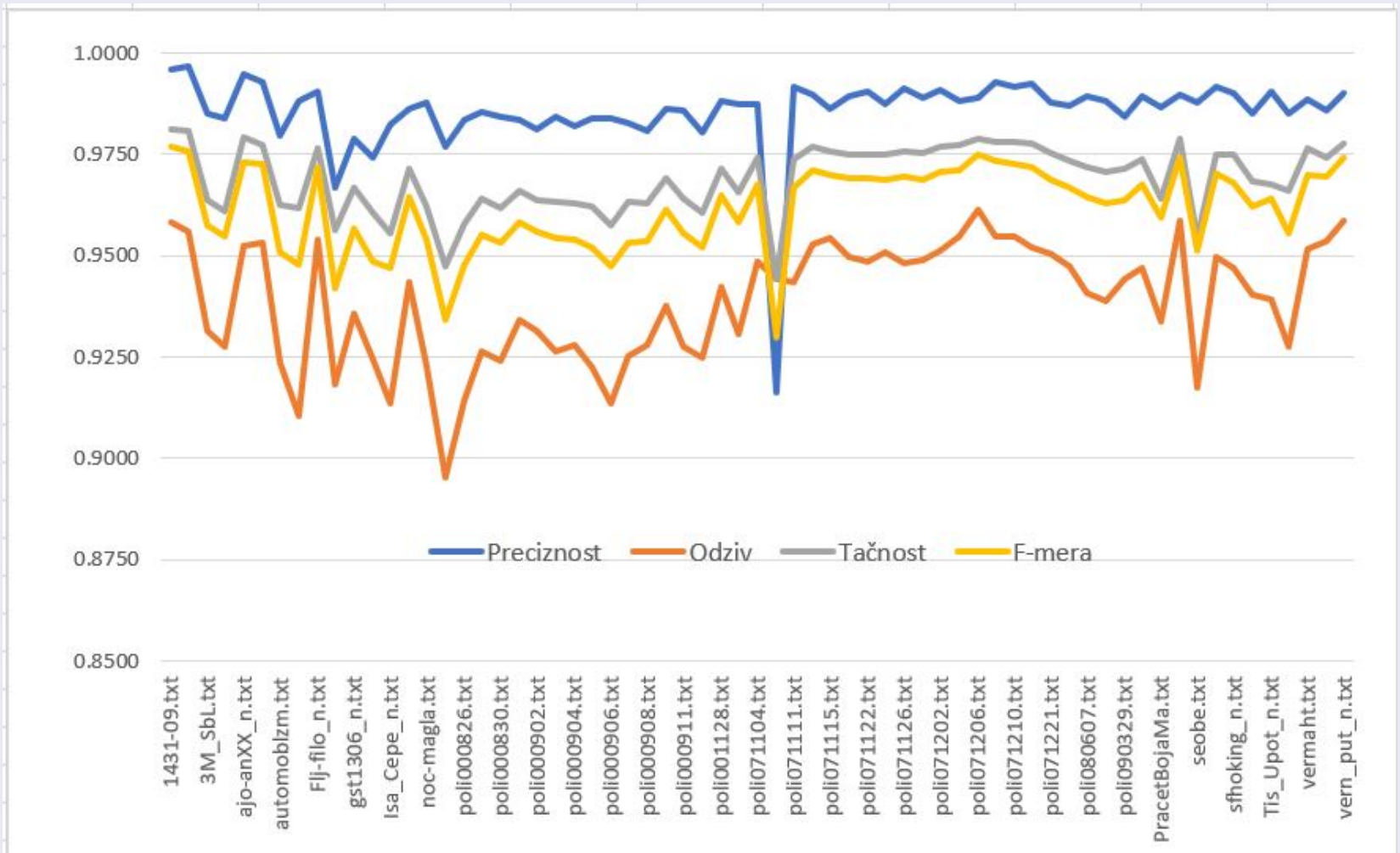
Najfrekvetniji tokeni i njhove ispravke – Žil Vern

Tip korekcije	Kandidati za promenu	Korigovani tekst	Frequency
6	*6(čaša(53)_časa(507))	časa	66
7	*7(što(36850)_sto(1268))	sto	41
7	*7(vas(1420)_vaš(363))	vas	40
7	*7(reci(237)_reči(2607)_reći(1448))	reči	28
7	*7(šta(8189)_sta(268))	sta	27
7	*7(Šta(8189)_Sta(268))	Sta	24
7	*7(Šer(10)_Ser(46)_Ser(171))	Ser	15
7	*7(preci(73)_preći(133))	preći	14
7	*7(Šiju(10)_Siju(24))	Siju	12
7	*7(uže(72)_uze(61))	uze	11
7	*7(Žar(35)_Zar(689))	Zar	11
7	*7(začelo(3)_zacelo(43))	zacelo	11
7	*7(žar(35)_zar(689))	zar	11
7	*7(nas(5623)_naš(3528))	nas	11
6	*6(čete(44)_ćete(485))	ćete	10
7	*7(Franciska(24)_Frančiška(1))	Franciska	10
7	*7(odluci(297)_odluči(314))	odluči	9
6	*6(žacu(1)_začu(35))	začu	8

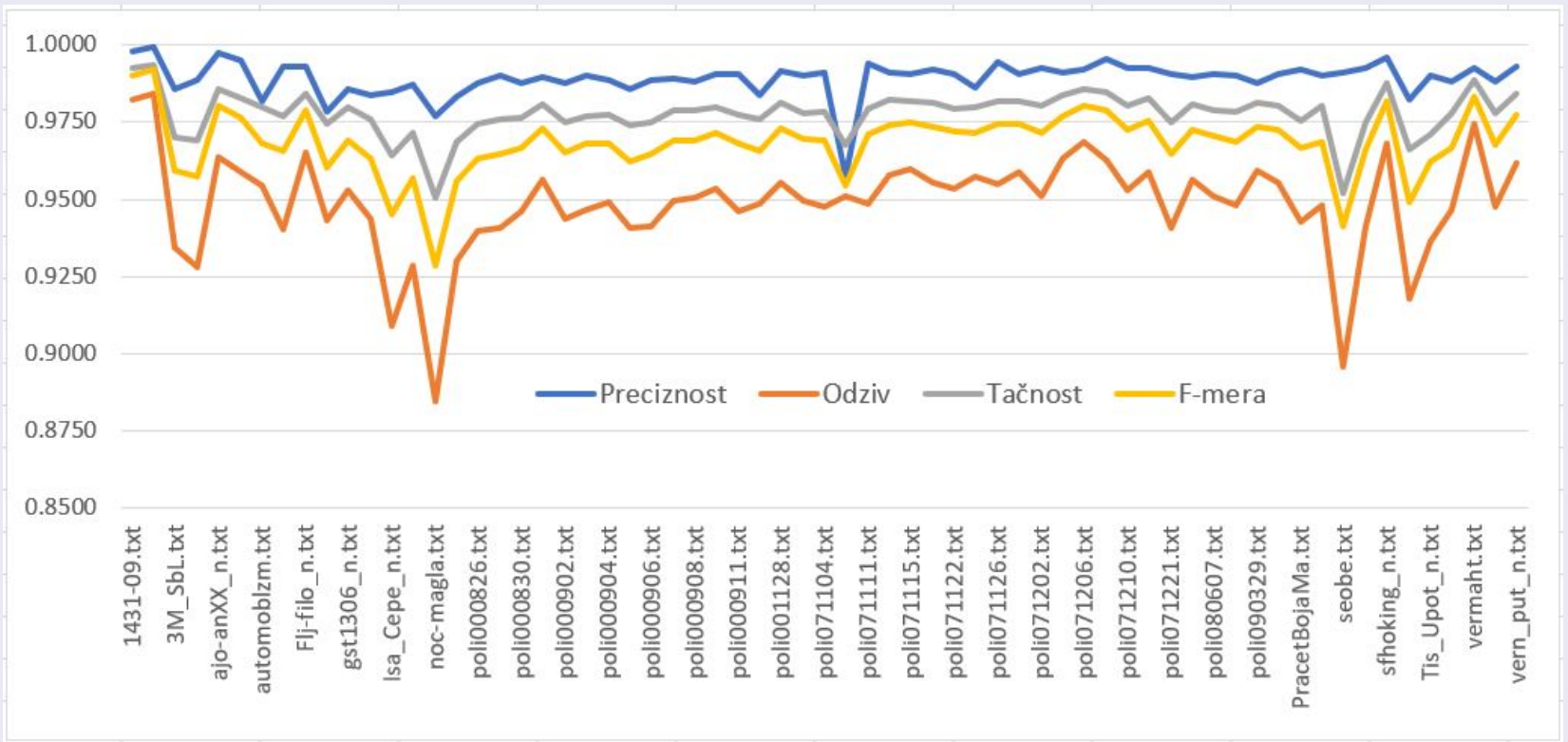
Primeri korekcija u kontekstu

Broj linije	Tip korekcij	Kandidati za promenu	Levi kontek	Korigovani teks	Desni kontekst
7	7	*7(privući(56)_privuci(1))	moglo na njega	privuci	pažnju.
34	7	*7(reci(237)_reči(2607)_reći(1448))	ni najbolje obavešteni nisu znali	reči	, a i gospodin Fog je bio poslednji
34	6	*6(ćuteci(1)_ćuteći(21))	ili velikodušnu stvar, on je davao	ćuteći	i neopaženo.
35	7	*7(mašti(33)_masti(93))	a su ljudi u svojoj nezadovoljenoj	masti	tražili još nešto iza toga.
36	7	*7(šta(8189)_sta(268))	talim putnicima; ukazivao je na to	sta	u stvari može biti istina, i njego
36	7	*7(reci(237)_reči(2607)_reći(1448))	ari može biti istina, i njegove su	reči	često bile proročke, jer ih je dog
37	6	*6(ćuteci(1)_ćuteći(21))	ranje vista.{S} U toj igri koja se	ćuteći	igrala, a koja je potpuno odgovara
37	7	*7(šume(224)_sume(152))	džep nego su predstavljali znatne	sume	u njegovom budžetu za dobročinstva
37	7	*7(reci(237)_reči(2607)_reći(1448))	za dobročinstva.{S} Treba,uostalom,	reči	da je gospodin Fog igrao radi igre
39	6	*6(vraćao(90)_vračao(1))	e pozivao za što nikakvog stranca,	vraćao	se kući samo na spavanje, tačno u
40	6	*6(čaša(53)_časa(507))	m članovima.{S} Od dvadeset četiri	časa	provodio je u svom stanu deset - u
40	7	*7(pića(297)_pica(15))	zila su na njegov sto sočna jela i	pica	iz klupskih kuhinja, ostava, ribar
40	6	*6(čašama(10)_časama(1))	{S} U klupskim kristalnim bocama i	čašama	bio je njegov "*7(šeri(13)_seri(3)
40	7	*7(šeri(13)_seri(3))	nim bocama i čašama bio je njegov "	seri	", njegov "porto", njegov "kleret"
40	7	*7(piće(254)_pice(12)_piče(1))	američkih jezera - održavao mu je	pice	u dovoljno svežem stanju.
44	7	*7(laksa(1)_lakša(26))	le, služba u njoj postajala je sve	lakša	.{S} Ipak je Fileas Fog zahtevao od
44	7	*7(što(36850)_sto(1268))	Forstera, - ovaj momak je pogrešio	sto	mu je za brijanje doneo vodu od os
51	7	*7(salona(88)_šalona(1))	enutku neko zakuca na vrata maloga	salona	u kome se
60	7	*7(vas(1420)_vaš(363))	ri džentlmen.{S} Preporučili su mi	vas	. -- Imam o vama dobra obaveštenja.

Distribucija parametara računatih po tokenima po dokumentima



Distribucija parametara računatih po frekvencijama po dokumentima



Poređenje sa rezultatima u radu Šantić et al. 2009

Table 3: Accuracy of diacritics restoration on different corpora

	Newspaper articles	Forum posts
Unrestored	80.72%	92.00%
Dictionary only (baseline)	97.07%	97.95%
Dictionary + Language model (unsmoothed)	97.65% (+ 0.6%)	98.38% (+0.4%)
Dictionary + Language model (WB smoothing)	98.81% (+1.8%)	99.35% (+1.4%)

- Oni su radili s tekstovima kojima nedostaju dijakritički znakovi, ali ne svi i ne uvek;
- Oni su pošli da je tekst na početku "tačan" 80,72% pa postaje sve tačniji, a kraju ostaje netačno 1,2%;
- Da bismo poredili, morali bismo da preračunamo naš račun na ovakav račun.

Poređenje sa rezultatima u radu Šantić et al. 2009

Table 3: Accuracy of diacritics restoration on different corpora

	Newspaper articles	Forum posts
Unrestored	80.72%	92.00%
Dictionary only (baseline)	97.07%	97.95%
Dictionary + Language model (unsmoothed)	97.65% (+ 0.6%)	98.38% (+0.4%)
Dictionary + Language model (WB smoothing)	98.81% (+1.8%)	99.35% (+1.4%)

- Oni su radili s tekstovima kojima nedostaju dijakritički znakovi, ali ne svi i ne uvek;
- Oni su pošli da je tekst na početku "tačan" 80,72% pa postaje sve tačniji, a kraju ostaje netačno 1,2%;
- Da bismo poredili, morali bismo da preračunamo naš račun na ovakav račun.

Poređenje sa rezultatima u radu Ljubešić et al. 2016

	wiki				tweet			
	P	R	FP	FN	P	R	FP	FN
hr	0.9901	0.9784	0.14%	0.30%	0.9831	0.9681	0.21%	0.41%
sr	0.9908	0.9705	0.12%	0.41%	0.9837	0.9642	0.26%	0.57%
sl	0.9852	0.9819	0.17%	0.21%	0.9745	0.9524	0.30%	0.58%

Table 3: Precision, recall, false positive and false negative error token percentage of the best performing system

- Mislim da su i oni automatski skidali dijakritičke znakove;
- Čini mi se da su oni računali parametre uzimajući u obzir sve tokene (*For Slovene standard text, approximately each 600th word will be erroneously rediacritised and each 500th word will fail to be rediacritised*);
- Da bismo poredili, morali bismo da preračunamo naš račun na ovakav račun (ali prvo da razumemo kako su računali).

Poređenje sa rezultatima u radu Ljubešić et al. 2016

	wiki				tweet			
	P	R	FP	FN	P	R	FP	FN
hr	0.9901	0.9784	0.14%	0.30%	0.9831	0.9681	0.21%	0.41%
sr	0.9908	0.9705	0.12%	0.41%	0.9837	0.9642	0.26%	0.57%
sl	0.9852	0.9819	0.17%	0.21%	0.9745	0.9524	0.30%	0.58%

Table 3: Precision, recall, false positive and false negative error token percentage of the best performing system

- Mislim da su i oni automatski skidali dijakritičke znakove;
- Čini mi se da su oni računali parametre uzimajući u obzir sve tokene (*For Slovene standard text, approximately each 600th word will be erroneously rediacritised and each 500th word will fail to be rediacritised*);
- Da bismo poredili, morali bismo da preračunamo naš račun na ovakav račun (ali prvo da razumemo kako su računali).

Ocena našeg sistema za vraćanje dijakritičkih znakova

Pozitivne strane

- Ima dobre strane sistema zasnovanog na pravilima (znanju), a to je transparentnost – svako ponuđeno rešenje ima logično objašnjenje, što olakšava ispravke;
- Samo unapređivanje leksičkih resursa će doprineti poboljšanju rada sistema (npr. duže i pouzdanije liste bigrama i trigrama, obuhvatniji rečnik polileksičkih jedinica);
- Modularnost, koraci se mogu lako izostavljati a grafovi kaskada zamenjivati.

Negative strane

- Utrošeno vreme za razvoj je sigurno veće nego da je korišćeno mašinsko učenje (ali je i zadovoljstvo istraživača veće!);
- Vreme izvršavanja nije zanemarljivo pa rešenje nije pogodno za interaktivne (npr. mobilne).

Ocena našeg sistema za vraćanje dijakritičkih znakova

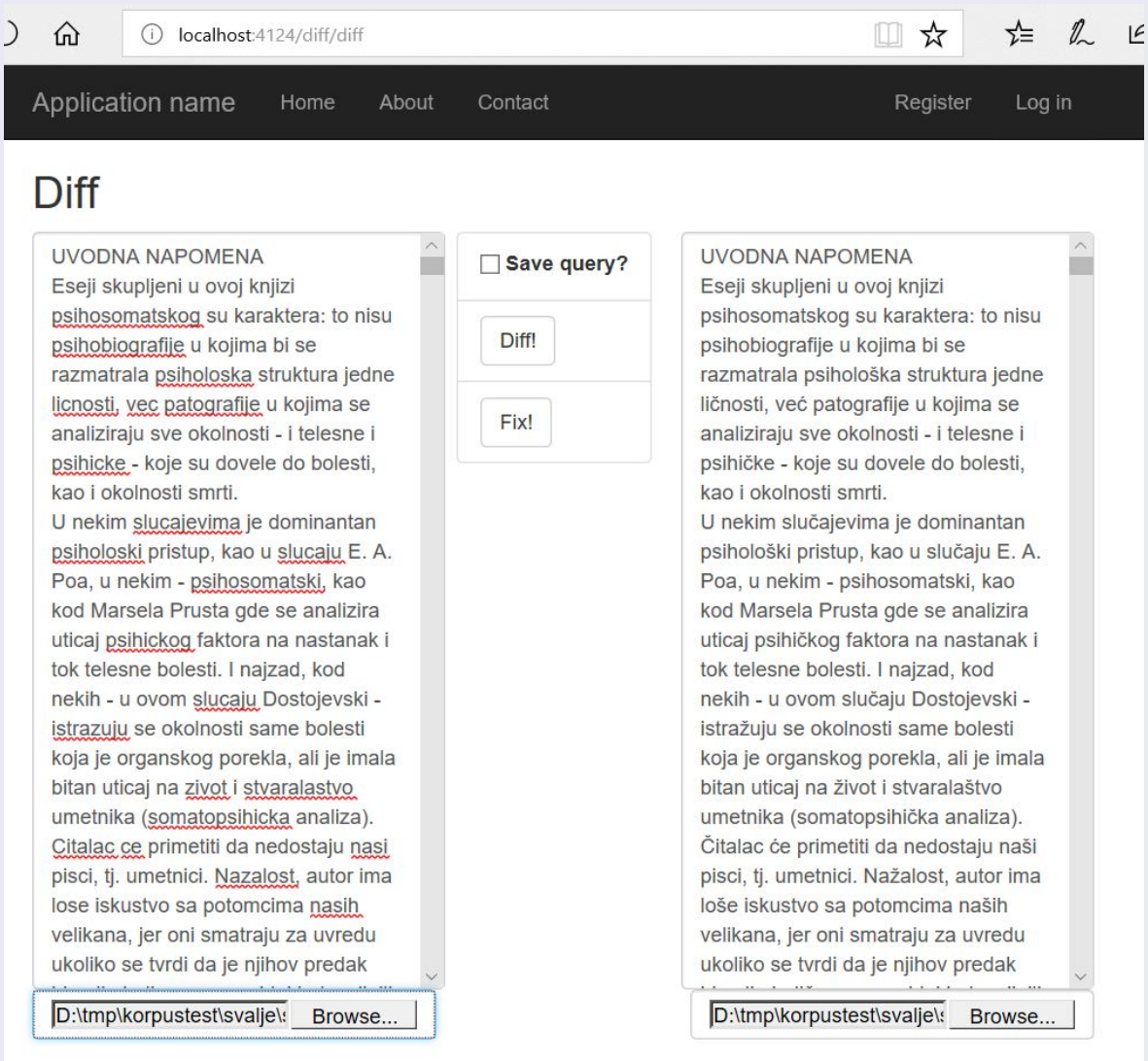
Pozitivne strane

- Ima dobre strane sistema zasnovanog na pravilima (znanju), a to je transparentnost – svako ponuđeno rešenje ima logično objašnjenje, što olakšava ispravke;
- Samo unapređivanje leksičkih resursa će doprineti poboljšanju rada sistema (npr. duže i pouzdanije liste bigrama i trigrama, obuhvatniji rečnik polileksičkih jedinica);
- Modularnost, koraci se mogu lako izostavljati a grafovi kaskada zamenjivati.

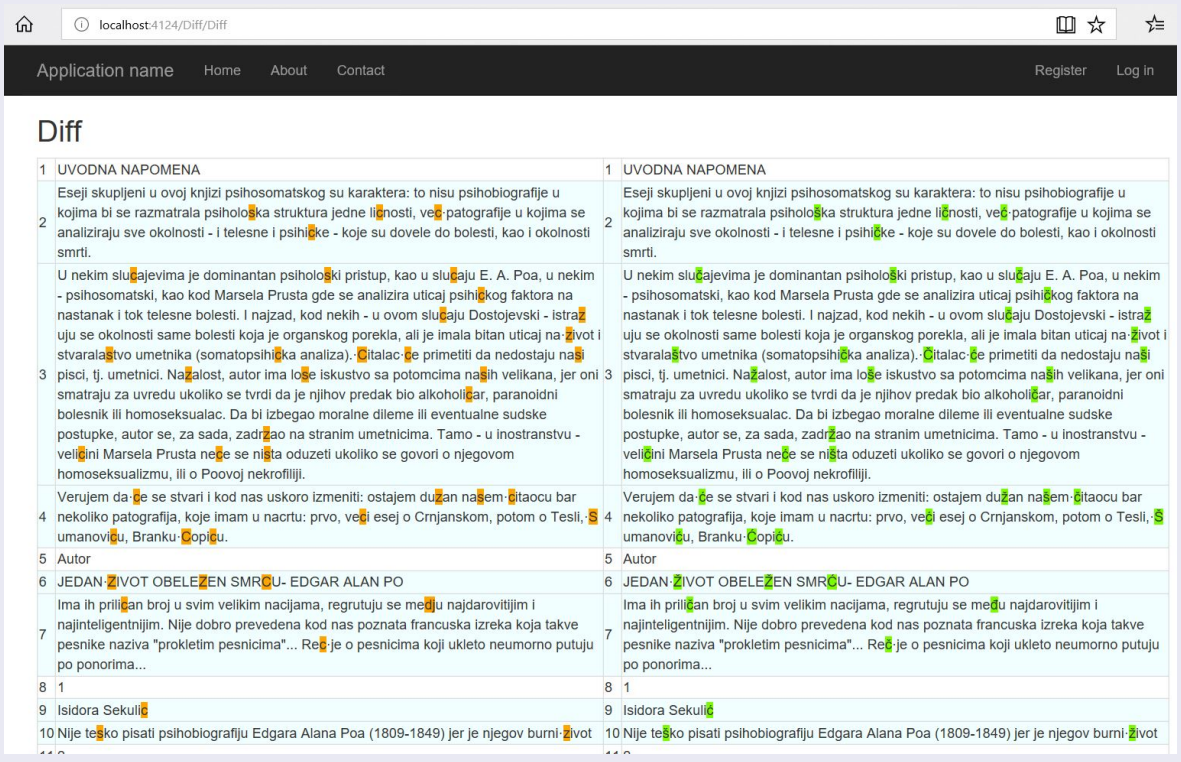
Negative strane

- Utrošeno vreme za razvoj je sigurno veće nego da je korišćeno mašinsko učenje (ali je i zadovoljstvo istraživača veće!);
- Vreme izvršavanja nije zanemarljivo pa rešenje nije pogodno za interaktivne (npr. mobilne).

Vraćanje dijakritičkih znakova na vebu – tiha ispravka



Vraćanje dijakritičkih znakova na vebu – glasna ispravka



1	UVODNA NAPOMENA	1	UVODNA NAPOMENA
2	Eseji skupljeni u ovoj knjizi psihosomatskog su karaktera: to nisu psihobiografije u kojima bi se razmatrala psihološka struktura jedne ličnosti, već patografije u kojima se analiziraju sve okolnosti - i telesne i psihičke - koje su dovele do bolesti, kao i okolnosti smrti.	2	Eseji skupljeni u ovoj knjizi psihosomatskog su karaktera: to nisu psihobiografije u kojima bi se razmatrala psihološka struktura jedne ličnosti, već patografije u kojima se analiziraju sve okolnosti - i telesne i psihičke - koje su dovele do bolesti, kao i okolnosti smrti.
3	U nekim slučajevima je dominantan psihološki pristup, kao u slučaju E. A. Poa, u nekim - psihosomatski, kao kod Marsela Prusta gde se analizira uticaj psihičkog faktora na nastanak i tok telesne bolesti. I najzad, kod nekih - u ovom slučaju Dostojevski - istraž uju se okolnosti same bolesti koja je organskog porekla, ali je imala bitan uticaj na život i stvaralaštvo umetnika (somatopsihička analiza). Čitalac će primetiti da nedostaju naši pisci, tj. umetnici. Nažalost, autor ima loše iskustvo sa potomcima naših velikana, jer oni smatraju za uvredu ukoliko se tvrdi da je njihov predak bio alkoholičar, paranoidni bolesnik ili homoseksualac. Da bi izbegao moralne dileme ili eventualne sudske postupke, autor se, za sada, zadržao na stranim umetnicima. Tamo - u inostranstvu - veličini Marsela Prusta neće se ništa oduzeti ukoliko se govori o njegovom homoseksualizmu, ili o Poovoj nekrofiliji.	3	U nekim slučajevima je dominantan psihološki pristup, kao u slučaju E. A. Poa, u nekim - psihosomatski, kao kod Marsela Prusta gde se analizira uticaj psihičkog faktora na nastanak i tok telesne bolesti. I najzad, kod nekih - u ovom slučaju Dostojevski - istraž uju se okolnosti same bolesti koja je organskog porekla, ali je imala bitan uticaj na život i stvaralaštvo umetnika (somatopsihička analiza). Čitalac će primetiti da nedostaju naši pisci, tj. umetnici. Nažalost, autor ima loše iskustvo sa potomcima naših velikana, jer oni smatraju za uvredu ukoliko se tvrdi da je njihov predak bio alkoholičar, paranoidni bolesnik ili homoseksualac. Da bi izbegao moralne dileme ili eventualne sudske postupke, autor se, za sada, zadržao na stranim umetnicima. Tamo - u inostranstvu - veličini Marsela Prusta neće se ništa oduzeti ukoliko se govori o njegovom homoseksualizmu, ili o Poovoj nekrofiliji.
4	Verujem da će se stvari i kod nas uskoro izmeniti: ostajem dužan našem čitaocu bar nekoliko patografija, koje imam u nacrtu: prvo, veći esej o Crnjanskom, potom o Tesli, Šumanoviću, Branku Copiću.	4	Verujem da će se stvari i kod nas uskoro izmeniti: ostajem dužan našem čitaocu bar nekoliko patografija, koje imam u nacrtu: prvo, veći esej o Crnjanskom, potom o Tesli, Šumanoviću, Branku Copiću.
5	Autor	5	Autor
6	JEDAN ŽIVOT OBELEŽEN SMRĆU- EDGAR ALAN PO	6	JEDAN ŽIVOT OBELEŽEN SMRĆU- EDGAR ALAN PO
7	Ima ih priličan broj u svim velikim nacijama, regrutuju se među najdarovitijim i najinteligentnijim. Nije dobro prevedena kod nas poznata francuska izreka koja takve pesnike naziva "prokletim pesnicima"... Reč je o pesnicima koji ukleto neumorno putuju po ponorima...	7	Ima ih priličan broj u svim velikim nacijama, regrutuju se među najdarovitijim i najinteligentnijim. Nije dobro prevedena kod nas poznata francuska izreka koja takve pesnike naziva "prokletim pesnicima"... Reč je o pesnicima koji ukleto neumorno putuju po ponorima...
8	1	8	1
9	Isidora Sekulić	9	Isidora Sekulić
10	Nije teško pisati psihobiografiju Edgara Alana Poa (1809-1849) jer je njegov burni život	10	Nije teško pisati psihobiografiju Edgara Alana Poa (1809-1849) jer je njegov burni život

Slično rešenje za slične probleme

Ispravka grešaka u OCR tekstu

- Ovo rešenje je slično po modularnosti i konsultovanju rečnika SMD – ispravljaju se samo reči koje nisu u rečniku rečima koje jesu;
- Ne postoji poseban rečnik za ispravke, jer vrste grešaka zavise od samog teksta, programa za OCR itd., ali tu pomaže modularnost – uključuju se grafovi koji su za dati tekst potrebni;
- Ovaj sistem još nije u potpunosti funkcionalan.

Pretvaranje iz ekavice u ijekavicu i obrnuto

- Sistem radi na sličan način kao sistem za vraćanje dijakritičkih znaka samo ima dva odvojena rečnika za "korekciju" za dva smera rada;
- I ovde se koriste frekvencije za izbor kandidata, a one su dobijene iz dva odvojena korpusa – ekavskog i ijekavskog;
- Ovaj sistem je funkcionalan ali nije obavljena evaluacija.

Slično rešenje za slične probleme

Ispravka grešaka u OCR tekstu

- Ovo rešenje je slično po modularnosti i konsultovanju rečnika SMD – ispravljaju se samo reči koje nisu u rečniku rečima koje jesu;
- Ne postoji poseban rečnik za ispravke, jer vrste grešaka zavise od samog teksta, programa za OCR itd., ali tu pomaže modularnost – uključuju se grafovi koji su za dati tekst potrebni;
- Ovaj sistem još nije u potpunosti funkcionalan.

Pretvaranje iz ekavice u ijekavicu i obrnuto

- Sistem radi na sličan način kao sistem za vraćanje dijakritičkih znaka samo ima dva odvojena rečnika za "korekciju" za dva smera rada;
- I ovde se koriste frekvencije za izbor kandidata, a one su dobijene iz dva odvojena korpusa – ekavskog i ijekavskog;
- Ovaj sistem je funkcionalan ali nije obavljena evaluacija.

Dve švalje nastavljaju da rade

Anna Ancher Ribareva žena šije 1890	Pavle Vasić Švalja 1948	Van Gogh Scheveningen 1882