(objavljeno u: Naučno-tehnički pregled, 1995, vol. 45, br. 6-7)

AUTOMATSKO KORIGOVANJE GREŠAKA NASTALIH OPTIČKIM ČITANJEM SRPSKOG TEKSTA

Pero Šipka
Filozofski fakultet u Novom Sadu
Odsek za psihologiju

Biljana Kosanović
Vojnotehnički institut, Beograd

Sažetak: Eksperimentalno je proverena efikasnost PAKoST-a, programa za postprocesiranje optički čitanog srpskog teksta prethodno razvijenog od istih autora. Uzorak je sadržao oko 70.000 reči latinično štampanog srpskog teksta, odabranih tako da ravnomerno obuhvate različite diskurse (naučni, književni i politički), fontove i kvalitete štampe. Tekst je optički pročitan pomoću Recognite Plus, jedinog uglednog komercijalnog OCR programa koji za sada podržava jugoslovenski set karaktera. Zatim je ASCII izlaz iz Recognite podvrgnut obradi pomoću PAKoST-a u režimu automatskog korigovanja.
Efikasnost postprocesiranja proverena je uz upotrebu dva algoritma: hibridnog kontekstnog postprocesora (HCP) koji je bio ugrađen u prethodnu verziju PAKoST-a i novog, složenijeg algoritma nazvanog MiniMax, implementiranog u najnoviju verziju programa. Efikasnost MiniMax-a u ispravljanju grešaka utvrđena je kako u odnosu na Recognitu, tako i u odnosu na HCP. Oba doprinosa tačnosti prepoznavanja testirana su statistički.
Testovi pokazuju da je novi algoritam bitno unapredio efikasnost PAKoST-a. Njegovom ugradnjom broj grešaka pri optičkom čitanju, izražen u rečima, smanjen je sa 7,90%, koliko proizvodi Recognita, na 4,39%. Pored toga MiniMax, za razliku od HCP-a, proizvodi snošljiv broj grešaka tipa II (grešaka koje sam proizvodi), što ohrabruje primenu PAKoST-a kao automatskog postprocesora.

1. Uvod

Za računarski podržani optički unos teksta može se reći da je definitivno osvojio birotehničku praksu. U najvećoj meri to se može zahvaliti proizvodnji relativno jeftinih stonih skanera koji takav unos omogućuju. Na svetskom tržištu umnožio se i broj jednako jeftinih programa za optičko prepoznavanje (Optical Character Recognition, OCR) čija je efikasnost sasvim zadovoljavajuća. Na taj način ispunjena su oba uslova da optički unos pokaže svoju nadmoć nad klasičnim manuelnim (daktilografskim) unosom, posebno onda kada se u računar unosi prethodno već štampani materijal. To podjednako važi za administrativno-poslovne, štamparsko-izdavačke, obrazovne i naučno-obrazovne namene računarskog teksta.
Sve to, na žalost, ne važi i za prostore našeg jezika. OCR programi, pa i oni najugledniji i najskuplji (Discover, WordScan, OmniPage) ne čitaju ćirilicu i po pravilu ne podržavaju obradu

naših znakova u latiničnom pismu (YuASCII). Kada to i čine (kao npr. Recognita Plus) njihova je efikasnost u prepoznavanju niska, što zbog nedovoljne brige za našeg korisnika, što zbog velike sličnosti tih grafema (međusobno ili naspram znakova internacionalnog skupa) i zato niske čitljivosti. Time je optički unos teksta za mnoge vidove primene u našim uslovima praktično invalidiran.
Glavna prepreka na putu ka efikasnom optičkom čitanju srpskog teksta lako je uočljiva. Algoritmi koji se koriste za optičko čitanje nužno se svojim značajnim delom temelje, ne samo na osobinama određenog pisma, već i određenog jezika. Naime, postupak optičkog čitanja uvek završava tzv. postprocesiranjem ili verifikacijom, kada se s osloncem na tekstualno okruženje overavaju hipoteze o verovatnom identitetu svakog provizorno pročitanog karaktera. Pošto svi algoritmi/metodi u prvom koraku čitaju sa znatnim brojem pogrešnih klasifikacija, u ovoj fazi se obavljaju preko potrebne ispravke. Tek s tim intervencijama efikasnost optičkog čitanja izdiže se na nivo koji ga čini (ekonomski) opravdanim.
Generalno se razlikuju tri metoda postprocesiranja, od kojih se jedan oslanja na rečnik, drugi na morfološke karakteristike jezika, a treći predstavlja njihovu kombinaciju.

(1) Metod oslanjanja na rečnik overava u prethodno oformljenoj listi reči (rečniku) legitimnost svake pročitane niske. Zahteva obimne rečnike i njihovo brzo pretraživanje. U te svrhe uspešno se koriste brojni algoritimi zasnovani na redundantnosti jezika (1, 2, 3, 4), raspodelama učestalosti leksičkih jedinica (5) i njihovom ekonomičnom indeksiranju (6). Ponuda zamene za neku nelegitimnu nisku (nereč, reč za koju se pretpostavlja da je pogrešno pročitana) utvrđuje se nekom merom sličnosti odnosno udaljenosti (7, 8, 9). Opšti nedostatak rečničkog pristupa je u velikim računarskim zahtevima u odnosu na obim rečnika i vreme obrade. To posebno važi za naš jezik.
(2) Statistički metod zasniva se na Bayes-ovom pristupu i tretmanu jezika kao Markov-ljevog procesa, a preuzet je iz kriptoanalize. Logička opravdanost ovog opšteg metoda je u jednostavnom saznanju da sve kombinacije susednih grafema (n-grama) u rečima nekog jezika nisu iste verovatnoće. Verovatnost svake od teorijskih kombinacija unutar jednog pisma i jezika utvrđuje se u okviru statističko-lingvističkih istraživanja i koristi za "predviđanje" svakog grafema u nisci u toku postprocesiranja. Nelegitimne ili malo verovatne kombinacije odbacuju se, kako bi se za njih ponudila odgovarajuća zamena. Najčešće se koriste tzv. Viterbi-jev (10), rekurzivni Bayes-ov algoritam (11) i tzv. probabilistička relaksacija (12). Većina tih algoritama neprikladna je za naš jezik. Opšta slabost statističkog pristupa je u tome što katkad nudi statističke konstrukte koji "zvuče" mogućno, ali ne postoje u datom jeziku.
(3) Hibridni metodi (13, 14) objedinjavaju prednosti dva osnovna pristupa i otklanjaju njihova nemala ograničenja. Opšte je iskustvo da oni daju najbolja rešenja u praksi.

Jedan takav hibridni kontekstni postprocesor (HCP) implementiran je i u prvu razvojnu verziju PAKoST-a - programa za automatsku korekciju srpskog teksta. HCP predstavlja nužnu adaptaciju algoritama razvijenih za engleski jezik. Njegova efikasnost provizorno je testirana tokom razvoja u više navrata na manjim uzorcima. Testovi su ukazivali na skromne, ali obećavajuće rezultate.
Noviji postupci postprocesiranja teže proširenju okruženja koje se koristi za overu legitimnosti ulaznog teksta. Kao osnovica za overu i korekciju sve više se koriste informacije o gramatičkoj frazeologiji (npr. tipičnim konstrukcijama i sintagmama) i pravopisnim pravilima, čime se statističko-lingvistička obrada uzdiže s leksičkog na sintaksički, pa i semantički nivo. Takođe se intenzivno istražuju i primenjuju razni modeli iz domena veštačke inteligencije i računarske

obrade prirodnog jezika (Natural Language Processing). Posebno se po efikasnosti ističu pristupi zasnovani na konsultovanju modula (ekspertnih sistema) koji se tokom obrade određenog ulaznog materijala "hrane" informacijama specifičnim za određenu vrstu i kvalitet štampe, ortografiju i diskurs. Time OCR programi dobijaju novi kvalitet, kome se s pravom pridaje atribut "inteligentnog prepoznavanja" (Intelligent Character Recognition, ICR). Poslednja verzija PAKoST-a predstavlja inovaciju na toj osnovi. Program sadrži kompleksniji algoritam (MiniMax) i nekoliko novih modula koji ga čine inteligentnim postprocesorom. Prvenstveno je namenjen obradi latiničnog štampanog teksta pročitanog komercijalnim OCR programima. PAKoST sadrži sledeće strukturalne elemente:

(1) provizorni rečnik srpskog jezika (80.000 reči/oblika),
(2) bazu podataka o susednosti grafema srpskog jezika,
(3) bazu podataka o tipografskim karakteristikama latiničnog štampanog teksta,
(4) bazu podataka o greškama tipičnim za optički unos,
(5) ortografski filtar,
(6) provizorni parser,
(7) centralni upravljački sistem i
(8) modul za odlučivanje.

Jednostavnim zahvatima na modulima 3. i 4. mogućno je proširiti primenu PAKoST-a na postprocesiranje optički pročitanog ćiriličnog teksta, optički pročitanog rukopisa, manuelno unetog teksta ili teksta unetog automatskim prepoznavanjem govora (Voice Recognition). Nezavisno od načina unosa PAKoST se može koristiti i kao overavač (spell-checker).

2. Cilj eksperimenta

U okviru ovog istraživanja eksperimentalno je ispitana efikasnost dva algoritma implementirana u dve verzije PAKoST-a. Istovremeno, rezultati su korišćeni za novo optimiziranje kriterijuma za izdvajanje ulaznih jedinica za automatsko korigovanje i druge vidove usavršavanja programa.

3. Metod

3.1. Uzorak

Kao uzorak za evaluaciju poslužio je prigodan (nameran, ali nepristrasan) izbor devet vrsta tekstova iz domena tri značajna diskursa srpskog jezika (naučnog, beletrističkog i političkog):

(1) Žika Lazić, Nepoznata svitanja, novela, 1 fragment, 5807 reči;
(2) Dragoslav Mihailović, Kad su cvetale tikve, roman, 1 fragment, 9264 reči;
(3) Aleksandar Tišma, Upotreba čoveka, roman, 1 fragment, 6397 reči;

(4) "Stanovništvo", časopis, godišta: 1988, 1991/92, 4 rada, 4 autora, bez tabela, 6962 reči;
(5) Miodrag Jovičić i dr (ur.), Izbori u uslovima višestranačkog sistema, fragmenti, 8985 reči;
(6) "Psihologija", časopis, godišta: 1980, 1988, 1993, 3 rada, 6 autora, bez tabela, 10262 reči;
(7) "Nedeljna Borba", 1992. godina, 9 priloga, 9 autora, bez antrfilea, 5841 reči;
(8) "Vreme", 1994. godina, 11 priloga, 15 autora, bez antrfilea, 7801 reči; i
(9) "Telegraf", 1994. godina, 7 priloga, 7 autora, bez antrfilea, 7078 reči.

Uzorak u celini obuhvatio je 68 397 reči, odnosno 436 294 znaka/grafeme. Svi tekstovi štampani su latiničnim pismom, različitim fontovima, među kojima je najčešće korišćen Times Roman. Kvalitet papira i otiska varirao je u širokom opsegu.

3.2. Postupak: čitanje i korekcija

Ulazni tekst za analizu dobijen je optičkim čitanjem pomoću stonog skanera Hewlett Packard ScanJet Plus. Skanerom je upravljano računarom PC klase s matičnom pločom Intel 80486 i SCSI kontrolerom. Za optičko čitanje korišćen je program Recognita Plus, verzija 2.0 xy, proizvođača SzKI[1]. Recognita je najugledniji evropski OCR program. Jedini je u klasi vrhunskog softvera koji bez "uvežbavanja" razlikuje YuASCII znakove. Program je zasnovan na algoritmu poznatom kao analiza kontura. Deklarisana tačnost prepoznavanja iznosi 99,99%, ali se ona ne postiže ni u idealnim uslovima. Komparativne prednosti Recognite su podrška velikog broja pisama i velika brzina, a nedostatak manja tačnost prepoznavanja. Recognita ima modul za interaktivno uvežbavanje, ali je njegova efikasnost sporna. Zato je eksperimentalno čitanje obavljeno na uobičajeni način.
Tekst pročitan Recognitom podvrgnut je korekciji pomoću obe pomenute verzije PAKoST-a, odnosno dva algoritma: HCP i MiniMax. U oba slučaja postupak korigovanja završavao je jednom od sledećih odluka:

(1) niska je sadržana u rečniku i prema tome predstavlja reč (overavam),
(2) ne postoji u rečniku i verovatno nije reč (ostavljam),
(3) ne postoji u rečniku, ali verovatno predstavlja reč (prihvatam) i
(4) ne postoji u rečniku, ali joj nalazim zamenu (zamenjujem/prepravljam).

Poput svih automatskih korektora PAKoST neizbežno donosi i pogrešne odluke. Postoje dve vrste pogrešnih odluka:

(1) nisku zamenjujem neodgovarajućom reči ili generišem nereč (kvarim) i
(2) prepravljam nereč u novu nereč (ne uspevam).

Uslovno se greškama mogu smatrati i sledeće klasifikacije:

(1) ostavljam neodgovarajuću reč (nepotrebno ostavljam) i
(2) prihvatam nereč (neumesno prihvatam)

Ove se pogrešne klasifikacije mogu smatrati uslovnima zato što PAKoST za njih obezbeđuje komfornu interaktivnu (neautomatsku) korekciju, tj. nudi listu reči-kandidata i uslove za

jednostavnu zamenu.

3.3. Obrada podataka

Obrada je izvršena pomoću programa Clipper jednostavnim upoređivanjem optički čitanih tekstova (uzorci) s datotekama koje su bile identične originalnim tekstovima (kriterijski tekstovi). Kriterijske datoteke formirane su, bilo korigovanjem uzoraka (sravnjivanjem s originalom), bilo preuzimanjem postojećih računarskih zapisa od izdavača, autora ili iz tzv. YU korpusa M. Henning-a[2].
U prvoj fazi obrade identifikovane su sve greške nastale pri optičkom čitanju. Izvršena je frekvencijska analiza i razvrstavanje pojedinačnih grešaka u kategorije: greške odbacivanja (neprepoznavanja, rejection errors), greške u čitanju YuASCII skupa, greške u čitanju znakova interpunkcije, greške u čitanju belina ("stapanje" i "rascepljivanje" reči) i greške u čitanju znakova međunarodnog skupa.
Zatim je skanirani tekst podvrgnut obradi pomoću programa PAKoST uz korišćenje oba testirana algoritma: hibridnog kontekstnog postprocesora (HCP) i modela MiniMax.
Efikasnost korigovanja, shvaćena kao razlika u tačnosti prepoznavanja između Recognite i svakog od dva algoritma PAKoST-a ponaosob, bila je izražena u rečima:

(1) proporcijom pravilnih odluka (tipa zamenjujem/prepravljam),
(2) zbirnom proporcijom pogrešnih odluka tipa neumesno prihvatam i tipa kvarim i
(3) proporcijom pogrešnih odluka tipa ne uspevam.

Doprinos PAKoST-a efikasnosti korigovanja, tj. značajnost razlika u proporciji grešaka između Recognite, HCP-a i MinMax-a ispitana je t-testom za značajnost razlika među proporcijama (15).

4. Rezultati

Rezultati (tabela 1. kolone 1 i 2) pokazuju da tačnost optičkog čitanja pomoću Recognite na svim uzorcima obilato odstupa od deklarisane, kao i one dobijene čitanjem engleskog teksta (16). Ipak, proporcija grešaka manja je od one utvrđene na nekim provizornim testovima u nas (17). Struktura grešaka Recognite inače je manje-više očekivana: najbrojnije su greške tipa odbacivanja (41.50%), a odmah zatim one u kojima učestvuju znakovi YuASCII (35.02%). Poseban problem čini visok procenat grešaka u kojima se reči "stapaju" ili "rascepljuju" (10.80%). Nije mali ni procenat grešaka u čitanju znakova interpunkcije (7.89%), mada se takve greške uglavnom smatraju bezbolnima.
Iz rezultata je takođe vidljivo da HCP predstavlja nepouzdanu osnovicu za korekciju. Na ukupnom uzorku on doduše značajno redukuje broj grešaka koje proizvodi Recognita, ali na nekim uzorcima njegova primena daje nezadovoljavajuće (Vreme) ili čak negativne efekte (Mihailović, Tišma). Nasuprot tome, MiniMax obezbeđuje pouzdanu korekciju na svim uzorcima, smanjujući broj grešaka Recognite na skoro polovinu (za 44,44%). Doprinos MiniMaxa ukupnoj tačnosti prepoznavanja statistički je značajan ne samo u odnosu na

Recognitu (t = 45.24, p = 0.001), već i u odnosu na HCP (t = 14.16, p = 0.001).

Tabela 1. Tačnost prepoznavanja pomoću Recognite i efikasnost korigovanja HCP-a i MiniMax-a

	Recognita		HCP		MiniMax	
	G_z	G_r	G_r	Dif	G_r	Dif
Lazić	1.78	8.25	5.96	27.77	4.77	42.17
Mihailović	1.37	4.17	5.04	-20.73	3.56	14.77
Tišma	1.47	4.85	6.10	-25.81	3.95	18.39
Stanovništvo	1.52	8.36	2.84	65.98	2.11	74.74
Jovičić (ur.)	0.82	4.80	3.55	25.99	2.78	42.00
Psihologija	2.27	12.02	8.02	33.31	6.78	43.60
N.Borba	2.51	12.26	6.21	49.30	5.68	53.63
Vreme	1.16	5.78	5.73	0.89	4.04	30.16
Telegraf	2.02	11.50	6.67	42.01	5.69	50.49
Ukupno	1.65	7.90	5.59	29.22	4.39	44.44

G_z- procenat grešaka u znacima G_r- procenat grešaka u rečima Dif - doprinos korekciji (u odnosu na Recognitu) izražen u procentima

Prednost MiniMax-a posebno je uočljiva iz rezultata (tabela 2) koji pokazuju strukturu i ispravnost odluka dva algoritma. Odluke se odnose na ukupan uzorak niski koje PAKoST ne sadrži u svom rečniku, pa zato podvrgava obradi pomoću testiranih algoritama. Iz kolona 4-10 vidi se da MiniMax postupa ispravnije u svim modalitetima odlučivanja. On ostvaruje uspešniju korekciju (zamenjujem i uspevam, kolona 8) i donosi manje pogrešnih odluka tipa zamenjujem i ne uspevam (kolona 10). Od posebne je važnosti da ukupnu proporciju grešaka tipa zamenjujem i kvarim (kolona 9) dramatično smanjuje. MiniMax je efikasan, i istovremeno nadmoćan u odnosu na HCP, i s obzirom na broj reči koje prihvata kao verovatne i ne podvrgava korekciji mada ne postoje u rečniku (prihvatam, kolona 5, odnosno 6). Pri svemu tome broj pogrešnih odluka tipa ostavljam (nereč, kolona 3) koje donosi MiniMax zanemarljivo je veći.

Tabela 2. Struktura grešaka s obzirom na modalitete odlučivanja*

		Ostavljam:			Prihvatam:			Zamenjujem i:			
		S	reč	nereč	S	reč	nereč	S	uspevam	kvarim	ne uspeva
		1	2	3	4	5	6	7	8	9	10
Lazić	HCP	43.48	54.91	45.09	14.13	90.70	9.30	42.39	59.95	25.32	14.73

	MiniMax	46.22	51.66	48.34	23.44	90.65	9.35	30.34	80.87	7.58	11.55
Mihailović	HCP	45.77	70.89	29.11	25.94	84.29	15.71	28.29	32.41	54.57	13.02
	MiniMax	46.16	70.29	29.71	36.21	87.23	12.77	17.63	57.78	32.44	9.78
Tišma	HCP	43.60	72.29	27.71	27.25	90.33	9.67	29.16	28.97	53.89	17.13
	MiniMax	44.50	70.82	29.18	40.05	92.29	7.71	15.44	55.29	21.76	22.94
Stanovništvo	HCP	19.46	66.24	33.76	33.91	96.85	3.15	46.63	81.51	13.73	4.75
	MiniMax	19.79	65.15	34.85	39.66	94.62	5.38	40.56	92.51	4.25	3.24
Jovičić (ur.)	HCP	29.05	60.59	39.41	34.45	76.80	23.20	36.50	58.88	25.44	15.68
	MiniMax	29.27	60.15	39.85	44.06	77.45	22.55	26.67	79.76	6.07	14.17
Psihologija	HCP	34.65	43.78	56.22	20.14	69.52	30.48	45.21	64.31	17.85	17.85
	MiniMax	35.87	42.29	57.71	27.60	73.71	26.29	36.53	79.86	4.72	15.42
N.Borba	HCP	31.43	39.20	60.80	15.52	78.75	21.25	53.06	75.87	11.15	12.98
	MiniMax	32.20	38.25	61.75	23.47	71.07	28.93	44.33	87.53	3.28	9.19
Vreme	HCP	39.60	65.32	34.68	31.59	77.33	22.67	28.81	41.81	40.35	17.84
	MiniMax	40.27	64.23	35.77	41.79	81.25	18.75	17.94	76.06	11.74	12.21
Telegraf	HCP	37.93	51.69	48.31	15.31	76.35	23.65	46.76	70.81	15.16	14.03
	MiniMax	38.69	50.68	49.32	22.55	77.26	22.74	38.76	84.05	3.50	12.45
Ukupno	HCP	36.05	58.07	41.93	23.99	82.19	17.81	39.96	61.12	24.78	14.10
	MiniMax	36.93	56.69	43.31	32.78	83.14	16.86	30.30	80.52	7.81	11.67
*sve vrednosti u tabeli izražene su u procentima											

U korekciji teksta koji je dobro "pokriven" rečnikom PAKoST-a, odnos ispravki i grešaka obe vrste može se smatrati sasvim zadovoljavajućim. Međutim, varijacije među uzorcima su znatne i ta odstupanja obavezuju na strože definisanje kriterijuma za izbor reči-kandidata u rečniku. Intervencija te vrste vodi smanjenju ukupnog broja uspešnih korekcija, ali zato smanjuje proporciju pogrešnih, što je poželjno za većinu praktičnih namena. Već na osnovu rezultata ovog istraživanja takva optimizacija je delimično obavljena.

5. Zaključak

Rezultati pokazuju da je automatska korekcija optički čitanog srpskog teksta pomoću domaćeg ICR softvera kakav je PAKoST ne samo mogućna, već i efikasna. Oslonjen na odgovarajući OCR, taj program izdiže čitljivost na nivo koji opravdava njegovu primenu, a time i komercijalizaciju.
Rezultati istovremeno upućuju na opšti zaključak da se zadovoljavajuće korigovanje grešaka optički čitanog srpskog teksta ne može postići, ni postojećim komercijalnim softverom, ni

primenom poznatih ICR algoritama koji su razvijeni za druge jezike. Takvo korigovanje mogućno je tek postupcima koji se zasnivaju na temeljnoj analizi morfoloških, leksičkih i sintaksičkih karakteristika srpskog jezika.

Literatura

(1) Fredkin, E., Trie memory, Commun. ACM, 3(9), 490-500 (1960)
(2) Knuth, D.E., Digital searching, The Art of Computer Programming, Vol. 3., 481-499, Addison-Wesley, Reading (1973)
(3) Blumer, A., J. Blumer, D. Haussler, A. Ehrenfeucht, M. T. Chen and J. Seiferas, The smallest automation recognising the subwords of a text, Theor. Comput. Sci., 40, 31-55 (1985)
(4) Appel, A.W. i G. J. Jacobson, The world's fastest scrabble program, Commun. ACM, 31(5), 572-579 (1988)
(5) Sheil, B.A., Miedian split trees: a fast lookup technique for frequently occurring keys, Commun. ACM, 21(11), 947-958 (1978)
(6) Kokonen, T. i E. Reuhkala, A very fast associative method for the recognition and correction of misspelt words, based on redundant hash addressing, Proc. 4th Int. Joint Conf. Pattern Recognition, 807-809 (1978)
(7) Okuda, T., E. Tanaka i T. Kasai, A method for the correction of garbled words based on the Levenshtein Metric, IEEE Trans. Comput. 25(2), 172-178 (1976).
(8) Masek, W.J. i M. S. Paterson, A faster algorithm computing string edit distance, J. Comput. System. Sci. 20(1), 18-3l (1980).
(9) Kashyap, R.L. i B. J. Oommen, Spelling correction using probabilistic methods, Pattern Recognition Lett., 2(4), 147-154 (1984)
(10) Viterbi, A.J., Error bounds for convolutional codes and an asymtopatically optimum decoding algorithm, IEEE Trans. Information Theory 13(2), 260-269 (1967)
(11) Shingal, R., Rosenberg, D. i G.T. Toussaint, A simplified heuristic version of a recursive Bayes algorithm for using context in text recognition, IEEE Trans. Systems Man Cybernetics 8 (5), 412- 414, (1975)
(12) Goshtasby, A. i R.W. Ehrich, Contextual word recognition using probabilistic relaxation labelling, Pattern Recognition 21(5), 455-462 (1988)
(13) Srihari, S.N., Hull, J.J. i R. Choudhuri, Integrating diverse knowledge sources in text recognition, ACM Trans. Office Information sys. 1(1), 68-87 (1983)
(14) Shinha, R.M.K. i B. Prasada, Visual text recognition throgh contextual processing, Pattern Recognition 21(5), 463-479 (1988)
(15) Petz, B., Osnovne statističke metode za nematematičare, SNL, Zagreb (1981)
(16) Seiter, C., OCR: the recognition you deserve, Macworld, 11, 92-7 (1993)
(17) Ristanović, D., Veština čitanja, Računari 75, 22-25 (1991)

AUTOMATIC ERROR CORRECTION IN OPTICAL CHARACTER RECOGNITION OF SERBIAN TEXT

Abstract: The effectiveness of the PAKoST, a program for postprocessing optically read Serbian text developed by the same authors, was experimentally tested. The sample consisted of about 70.000 words of the Serbian Latin printed text compiled from nine sources covering evenly three discourses (scientific, literary, and political), different fonts, and various printing

qualities. The text was optically read by Recognita Plus, the only prestigious commercial OCR software that at present support Yugoslav character set. The ASCII Recognita's output was then processed by PAKoST in the automatic correction mode.
The effectiveness of the postprocessing was checked by using two algorithms: a hybrid contextual postprocessor (HCP) that was build into the previous version of PAKoST and a new more complex algorithm called MiniMax implemented in the last version of the program. Both general correctability (against Recognita) and incremental correctability (against HCP) of MiniMax was calculated and statistically tested.
The tests demonstrated that the new algorithm brought to PAKoST substantial improvement in correctability. It reduced Recognita word error rate from 7.90% to 4,39%. Furthermore, MiniMax, unlike HCP, produces a tolerable amount of type II errors (new errors of its own), encouraging the use of PAKoST as an automatic postprocessor.

Fusnote

[1] U okviru eksperimenta testirani su i programi ReadRight i Recognita 1.1., ali se ti rezultati ovde ne prikazuju; oba programa imaju toliko nisku tačnost prepoznavanja da rezultati ne zaslužuju pažnju.

[2] Zahvaljujemo se Naučnoj knjizi, BIGZ-u, Stanovništvu, Stanislavu Fajgelju, dr Branku Pokrajcu i Dušanki Vukićević na ustupljnim datotekama.