Poboljšanje poluautomatske morfološke anotacije primenom transformacionih pravila

Aleksandar D. Kupusinac, Milan S. Sečujski, *Member, IEEE*

***Sadržaj* — U ovom radu predstavljena je mogućnost poboljšanja tačnosti algoritma za poluautomatsku morfološku anotaciju tekstova na srpskom jeziku primenom transformacionih pravila. Transformaciona pravila su dobijena metodom automatskog učenja na obučavajućem skupu koji je prethodno anotiran primenom algoritma za poluautomatsku anotaciju. U radu je ukazano na značaj rešavanja ovog problema i na pojedine probleme karakteristične za morfološku anotaciju tekstova na jezicima sa bogatom morfologijom, kao što je srpski jezik.**

***Ključne reči* — govorne tehnologije, morfološka anotacija, obrada prirodnog jezika, transformaciona pravila.**

I. UVOD

MORFOLOŠKA anotacija zauzima značajno mesto u skoro svim aplikacijama jezičkih tehnologija, kao što su automatsko prevođenje testova i automatsko izvođenje zaključaka na osnovu teksta, a pogotovo govornih tehnologija – automatsko prepoznavanje i sinteza govora na osnovu teksta. Morfološka anotacija teksta podrazumeva određivanje gramatičkog statusa svake reči u tekstu (vrste reči i vrednosti odgovarajućih morfoloških kategorija) i kao takva predstavlja problem izuzetno zavisan od jezika. Složenost morfologije jezika sa bogatom morfologijom (složene morfološke kategorije, izvođenje reči upotrebom sufiksa i prefiksa, relativno slobodan red reči u rečenici i sl.) dovodi do poznatog problema manjka podataka, stoga su za istraživanje i razvoj algoritama za morfološku anotaciju tekstova na jezicima sa bogatom morfologijom neophodne veoma obimne baze ručno anotiranog teksta. Da bi se postigla približno ista tačnost, obuka algoritama za morfološku anotaciju tekstova na jezicima sa bogatom morfologijom zahteva znatno veću količinu ručno anotiranih tekstova od obuke algoritama za morfološku anotaciju tekstova na jezicima sa siromašnom morfologijom. Manjak podataka predstavlja nezaobilazan problem za morfološku anotaciju tekstova na svim jezicima, a pogotovo na jezicima sa bogatom morfologijom, a uz to razvoj ručno anotiranih tekstualnih baza predstavlja veoma skup i vremenski zahtevan posao.

U okviru projekta "AlfaNum" na Fakultetu tehničkih nauka u Novom Sadu razvijen je ekspertski algoritam, koji sa velikom tačnošću obavlja poluautomatsku morfološku anotaciju tekstova na srpskom jeziku. Pored ovog algoritma, na Fakultetu tehničkih nauka u Novom Sadu razvijen je i algoritam za potpuno automatsku morfološku anotaciju tekstova na srpskom jeziku koji se bazira na sekvencijalnoj primeni transformacionih pravila. Kombinacijom ova dva algoritma mogu se poboljšati performanse algoritma za poluautomatsku morfološku anotaciju tekstova na srpskom jeziku.

Ovaj rad je delimično finansiran od strane Ministarstva nauke i zaštite životne sredine Republike Srbije, u okviru projekta "Razvoj govornih tehnologija na srpskom jeziku i njihova primena u 'Telekomu Srbija'" (TR-6144A).

A. D. Kupusinac, Fakultet tehničkih nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-4852417; faks: 381-21-6350727; e-mail: sasak@uns.ns.ac.yu).

M. S. Sečujski, Fakultet tehničkih nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija; (e-mail: secujski@uns.ns.ac.yu).

II. O SRPSKOM JEZIKU

Srpski jezik spada u indoevropske, južnoslovenske jezike, a govori ga oko 11 miliona ljudi u Srbiji zajedno sa dijasporom [1]. Zbog složenih morfoloških kategorija, izvođenja reči upotrebom sufiksa i prefiksa, relativno slobodnog reda reči u rečenici i sl. srpski jezik, kao i drugi slovenski jezici, spada u grupu jezika sa bogatom morfologijom. Na primer, kompletna deklinacija prideva obuhvata sedam padeža, tri roda i dva broja, uključujući supletivnu promenu za množinu srednjeg roda.

Složenost morfologije srpskog jezika dovodi do poznatog problema manjka podataka. Postojeće tekstualne baze na srpskom jeziku najvećim delom nisu morfološki anotirane. Izuzetke predstavljaju korpus srpskog jezika razvijen na Institutu za eksperimentalnu fonetiku i patologiju govora u Beogradu, koji obuhvata tekstove iz perioda od 12. do 20. veka i u kome se nalazi oko 11 miliona ručno anotiranih reči [2], zatim, srpski prevod romana "1984." od Džordža Orvela sa oko 100 hiljada ručno anotiranih reči, koji predstavlja deo MULTEXT-East jezičkih resursa za srpski jezik [3] i AlfaNum korpus srpskog jezika sa oko 100 hiljada reči, koje su prethodno anotirane pomoću algoritma za poluautomatsku morfološku anotaciju razvijenog na Fakultetu tehničkih nauka u Novom Sadu [4,5], posle čega su greške ručno ispravljene. Od pomenutih korpusa jedino AlfaNum korpus srpskog jezika sadrži informaciju vezanu za tip i poziciju akcenta, koja je od izuzetne važnosti za primenu korpusa u aplikacijama govornih tehnologija.

III. OPIS ALGORITMA

A. Poluautomatska morfološka anotacija

Cilj algoritma za morfološku anotaciju je da za svaku reč u tekstu sa što većom tačnošću identifikuje o kojoj se

vrsti reči radi, kao i koje su konkretne vrednosti njenih morfoloških kategorija. Na primer, za imenicu je potrebno odrediti rod, broj i padež. Kompletna informacija o vrsti reči i vrednostima morfoloških kategorija obuhvaćena je jedinstvenom oznakom (engl. *tag*), tako da se problem morfološke anotacije svodi na dodelu oznake svakoj reči iz skupa mogućih oznaka za datu reč. Na primer, reč *gomila* može biti imenica ženskog roda u nominativu jednine, kao i genitivu množine, a može biti i glagol u trećem licu jednine prezenta. Dok u jezicima sa siromašnom morfologijom oznaka sadrži samo informaciju o vrsti reči, u jezicima sa bogatom morfologijom oznaka sadrži daleko više informacija.

Algoritam za poluautomatsku morfološku anotaciju [4] može se podeliti u nekoliko procedura. Nakon početnog rastavljanja ulaznog teksta na nizove znakova razdvojene razmacima i interpunkcijom – *tokene*, vrši se upit u morfološki rečnik [6] i za svaku reč se sastavlja lista mogućih oznaka. Sledeći korak je analiza konteksta, koja razmatra reč u kontekstu i pokušava da joj dodeli oznaku na osnovu mogućih reči u njenoj bližoj okolini. Ulazni podaci za konteksnu analizu sastoje se od liste mogućih oznaka za sve reči u rečenici. Neka je potrebno anotirati rečenicu $W=w_1w_2...w_N$. Svaka od reči w_i ima odgovarajući skup oznaka $T_i=\{t_{ij} \mid j=1, 2,..., N_i\}$ i njena stvarna oznaka t_i biće jedna od t_{ij}, $j=1, 2,..., N_i$. U inicijalnom koraku razmatraju se hipoteze dužine 1, koje sadrže samo prvu reč u rečenici $H_1=\{(t_{1j}) \mid j=1, 2,..., N_1\}$. U svakom sledećem koraku algoritma, svaka mogućnost za sledeću reč kombinuje se sa svakom od postojećih parcijalnih hipoteza. Skup svih parcijalnih hipoteza dužine 2 je $H_2=\{(t_{1m}, t_{2n}) \mid m=1, 2,..., N_1, \ n=1, 2,..., N_2\}$. Svaki put kada se doda nova reč, svakoj od parcijalnih hipoteza dodeljuju se bodovi, u skladu sa verovatnoćom da reč sa odgovarajućom oznakom može da usledi u datom kontekstu. Kriterijumi za određivanje broja poena koji će biti dodeljen hipotezama zasnovani su na pravilima definisanim na osnovu statistika pojedinih vrsta reči u srpskom jeziku, kao i određenih zavisnosti između njih. Definisana su i dodatna pravila koja se odnose na odgovarajuće zavisnosti u slučaju konkretnih reči. Neki od primera obrazaca za pravila opšteg tipa su sledeći:

Dodeliti n bodova parcijalnoj hipotezi $h=(w_1, w_2,..., w_l)$:

- ako reč w_l ima oznaku t_i,
- ako reč w_l ima oznaku t_i, a reč w_{l-1} ima oznaku t_j,
- ako reč w_k ima oznaku t_i, reč w_{l-1} ima oznaku t_j, a w_{l-2} ima oznaku t_k ,
- ako reč w_l ima ima oznaku t_i, reč w_{l-1} ima oznaku t_j, a vrednost morfološke kategorije c sadržane u oznaci t_i je jednaka (nije jednaka) vrednost odgovarajuće morfološke kategorije sadržane u oznaci t_j ,
- ako reč w_l ima oznaku t_i, reč w_{l-1} ima oznaku t_j i sve vrednosti morfoloških kategorija $c_1, c_2,..., c_k$ sadržanih u oznaci t_i su jednake (nisu jednake) vrednostima odgovarajućih morfoloških kategorija sadržanih u oznaci t_j.

Ako broj parcijalnih hipoteza prevaziđe unapred zadato ograničenje L, samo će L hipoteza sa najvećim brojem bodova biti zadržano, a sve ostale biće odbačene. Procedura se nastavlja dok se sve reči ne uključe, a tada se hipoteza sa najvećim prikupljenim brojem bodova uzima za procenu stvarnog niza oznaka $T=t_1t_2...t_N$. Dakle, kao izlaz algoritma dobija se spisak reči sa (verovatno) tačnim oznakama, kao i odgovarajuća akcentuacija celokupne rečenice.

B. Transformaciona pravila

Algoritam za automatsku morfološku anotaciju zasnovan na transformacionim pravilima [7] obuhvata nekoliko faza: faza inicijalne anotacije, faza dobijanja transformacionih pravila i faza testiranja. Korpus se deli na: skup za obuku i skup za test. U fazi inicijalne anotacije posmatra se relativna učestanost svake oznake u skupu za obuku. Pošto je svaka reč već unapred ručno anotirana, poznata je tačna oznaka svake reči. Na osnovu upita u morfološki rečnik, za svaku reč može se dobiti skup mogućih oznaka. Druga faza obuhvata dobijanje transformacionih pravila. Neka je potrebno anotirati rečenicu $W=w_1w_2...w_N$. Transformaciona pravila se dobijaju potpuno automatski na osnovu sledećih šablona:

Posmatra se reč w_i. Potrebno je promeniti njenu oznaku t_i (dobijenu u inicijalnoj anotaciji) u oznaku t_j :

- ako prethodna (sledeća) reč ima oznaku t_k,
- ako reč dva mesta unazad (unapred) ima oznaku t_k,
- ako bilo koja od dve prethodne (sledeće) reči ima oznaku t_k,
- ako bilo koja od tri prethodne (sledeće) reči ima oznaku t_k,
- ako prethodna reč ima oznaku t_k, a sledeća reč ima oznaku t_l,
- ako prethodne (sledeće) dve reči imaju oznake t_k i t_l,
- ako je prethodna (sledeća) reč L,
- ako su prethodne (sledeće) dve reči L_1 i L_2.

Ukupna korist od svakog transformacionog pravila dobija se kao razlika broja ispravljenih i izazvanih grešaka. Na taj način, za svaki šablon, moguće je izdvojiti pouzdana pravila koja u najvećoj meri podižu tačnost anotacije. Zatim sledi faza testiranja, u kojoj se izabrana, dovoljno pouzdana pravila primenjuju na skup za testiranje. Skup za testiranje prvo prolazi kroz inicijalnu anotaciju, a zatim se za svaku reč traži odgovarajuće transformaciono pravilo koje se odnosi na njenu oznaku, odgovara kontekstu u kome se data reč nalazi i ima što veću korist. Nađeno pravilo će biti primenjeno i za datu reč će doći do zamene oznaka u skladu sa nađenim pravilom .

C. Poboljšanje poluautomatske morfološke anotacije

Tačnost algoritma za poluautomatsku morfološku anotaciju se može povećati kombinovanjem sa algoritmom zasnovanim na transformacionim pravilima. Faza testiranja algoritma zasnovanog na transformacionim pravilima može se modifikovati tako da se ne posmatra relativna učestanost svake oznake u skupu za obuku, već da se za svaku reč prihvata oznaka dobijena algoritmom za poluautomatsku morfološku anotaciju kao inicijalna oznaka za datu reč, posle čega se primenjuju dobijena trans-

formaciona pravila. Za svaku reč traži se odgovarajuće transformaciono pravilo koje se odnosi na datu reč, odgovara kontekstu u kome se data reč nalazi i ima što veću korist. Nađeno pravilo će biti primenjeno i za datu reč će doći do zamene oznaka u skladu sa nađenim pravilom.

IV. Eksperiment i rezultati

Za obuku i testiranje algoritama opisanih u ovom radu korišćen je AlfaNum korpus srpskog jezika sa oko 100 hiljada reči. U tabeli 1 su prikazane uporedni vrednosti greške morfološke anotacije za sva tri algoritma. Testiranje algoritma za poluautomatsku morfološku anotaciju (u tabeli je označen kao PMA) na test skupu od 100 hiljada reči je pokazalo da greška morfološke anotacije iznosi 12.28%. Testiranje algoritma za automatsku morfološku anotaciju zasnovanog na transformacionim pravilima (u tabeli je označeno kao AMTR) pri čemu je inicijalna anotacija izvedena tako što je ceo korpus od oko 100 hiljada reči podeljen na dva dela: skup za obuku (oko 80 hiljada reči) i skup za testiranje (oko 20 hiljada reči) i dobijena greška morfološke anotacije iznosi 12.71% [8].

Da bi se video efekat poboljšanja poluautomatske morfološke anotacije primenom transformacionih pravila, odnosno šta se dobija ako se inicijalna automatska morfološka anotacija zameni poluautomatskom morfološkom anotacijom (u tabeli je označeno kao PMA-AMTR) korpus je podeljen u dva dela: skup za obuku (oko 80 hiljada reči) i skup za testiranje (oko 20 hiljada reči). Skup za obuku je prvo anotiran pomoću algoritma za poluautomatsku morfološku anotaciju, a zatim su formirana transformaciona pravila. Testiranje je izvedeno tako što je skup za testiranje prvo anotiran pomoću algoritma za poluautomatsku morfološku anotaciju, a zatim su primenjena transformaciona pravila i greška morfološke anotacije iznosi 9.99%. Dakle, opisano poboljšanje algoritma za poluautomatsku morfološku anotaciju primenom transformacionih pravila je smanjilo grešku morfološke anotacije sa 12.28% na 9.99%.

Tabela 1: Rezultati algoritama.

Algoritam	***Greška morfološke anotacije***
PMA	12.28%
AMTR	12.71%
PMA-AMTR	9.99%

Treba napomenuti da od veličine skupa oznaka zavisi i tačnost anotacije i očigledno važi da se sa manjim brojem oznaka može očekivati veća tačnost. Međutim, skup oznaka ipak mora biti dovoljno velik da se ne bi izgubila informacija koja je od suštinske važnosti za primenu dobijene morfološke anotacije. Prema tome, veličina skupa oznaka predstavlja kompromis između tačnosti i detaljnosti. Veličina skupa oznaka u jezicima sa siromašnijom morfologijom (npr. engleski jezik) tipično se kreće u opsegu od 40 do 70 različitih oznaka. Međutim, u jezicima sa bogatom morfologijom (npr. srpski, češki, mađarski, rumunski itd.) veličina test skupa se kreće u opsegu od 500 pa i preko 1000 različitih oznaka.

V. Zaključak

U ovom radu opisani su pojedini algoritmi za morfološku anotaciju i dati uporedni rezultati greške morfološke anotacije u okviru eksperimenta na korpusu sa oko 100 hiljada reči. Takođe, u radu je predstavljena mogućnost poboljšanja tačnosti algoritma za poluautomatsku morfološku anotaciju tekstova na srpskom jeziku primenom transformacionih pravila, usled čega se greška morfološke anotacije smanjila sa 12.28% na 9.99%. Kao dalji pravci istraživanja nameću se istraživanje novih metoda kombinovanja poznatih algoritama morfološke anotacije, kao i istraživanje mogućnosti dobijanja što opštijih pravila, što bi bilo od velikog značaja posebno za jezike sa bogatom morfologijom.

Literatura

[1] B. F. Grimes, *Ethnologue – languages of of the world.* SIL International, 1996.

[2] Đ. Kostić, *Kvantitativan opis strukture srpskog jezika: korpus srpskog jezika.* Institut za eksperimentalnu fonetiku i patologiju govora, Filozofski fakultet, Beograd, 2001.

[3] C. Krstev, D. Vitas, T. Erjavec, "MULTEXT-East resources for Serbian," *IS-LTC 2004*, Ljubljana, pp. 108–114.

[4] M. Sečujski, V. Delić, "A Software Tool for Semi-Automatic Part-of-Speech Tagging and Sentence Accentuation in Serbian Language," in *Proc.IS-LTC 2006*, Ljubljana, pp. 226-229.

[5] M. Sečujski, "Softver za izradu korpusa morfološki anotiranih tekstova na srpskom jeziku," *DOGS 2006*, Vršac, pp. 30-33.

[6] M. Sečujski, "Akcenatski rečnik srpskog jezika namenjen sintezi govora na osnovu teksta," *DOGS 2002*, Bečej, pp. 17-20.

[7] E. Brill, "A Simple Rule-Based Part-of-Speech Tagger," *ANLP 1992*, Trento, pp. 152-155.

[8] A. Kupusinac, M. Sečujski, "An Algorithm for Part-of-Speech Tagging in Serbian Language," *ISIRR 2007*, Novi Sad

Abstract

In this paper we have analysed a strategy aimed at improving accuracy of an algorithm for semi-automatic morphological analysis (part-of-speech tagging) of texts in Serbian language, based on application of transformation rules. The transformation rules have been obtained by automatic learning procedures on a dataset previously annotated using an algorithm for semi-automatic morphological analysis based on hand-written rules. The paper discusses the importance of solving this problem, especially in view of all the particularities of morphologically rich languages such as Serbian.

AN IMPROVEMENT OF SEMI-AUTOMATIC POS-TAGGING BY TRANSFORMATION RULES

Aleksandar D. Kupusinac, Milan S. Sečujski