UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA

DOKTORSKA DISERTACIJA

STEVAN OSTROGONAC

NOVI SAD, 2018

UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
DEPARTMAN ZA ENERGETIKU, ELEKTRONIKU I TELEKOMUNIKACIJE
KATEDRA ZA TELEKOMUNIKACIJE I OBRADU SIGNALA

DOKTORSKA DISERTACIJA:

MODELI SRPSKOG JEZIKA I NJIHOVA PRIMENA U GOVORNIM I JEZIČKIM TEHNOLOGIJAMA

Mentor:
Prof. dr Milan Sečujski

Kandidat:
Stevan Ostrogonac

NOVI SAD, 2018

UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: **RBR**	
Identifikacioni broj: **IBR**	
Tip dokumentacije: **TD**	Monografska dokumentacija
Tip zapisa: **TZ**	Tekstualni štampani materijal
Vrsta rada (dipl., mag., dokt.): **VR**	Doktorska disertacija
Ime i prezime autora: **AU**	Stevan Ostrogonac
Mentor (titula, ime, prezime, zvanje): **MN**	prof. dr Milan Sečujski
Naslov rada: **NR**	Modeli srpskog jezika i njihova primena u govornim i jezičkim tehnologijama
Jezik publikacije: **JP**	srpski
Jezik izvoda: **JI**	srpski/engleski
Zemlja publikovanja: **ZP**	Srbija
Uže geografsko područje: **UGP**	Vojvodina
Godina: **GO**	2018.
Izdavač: **IZ**	Autorski reprint
Mesto i adresa: **MA**	FTN, Trg Dositeja Obradovića 6, 21000 Novi Sad
Fizički opis rada: **FO**	6 poglavlja / 98 stranica / 0 citata /15 slika / 10 grafikona / 19 tabela / 89

	referenci / 5 priloga
Naučna oblast: **NO**	Elektrotehničko i računarsko inženjerstvo
Naučna disciplina: **ND**	Telekomunikacije i obrada signala
Predmetna odrednica, ključne reči: **PO**	Obrada prirodnog jezika Računarska lingvistika
UDK	
Čuva se: **ČU**	u biblioteci Fakulteta tehničkih nauka u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad
Važna napomena: **VN**	
Izvod: **IZ**	*Statistički jezički model, u teoriji, predstavlja raspodelu verovatnoća nad skupom svih mogućih sekvenci reči nekog jezika. U praksi, to je mehanizam kojim se estimiraju verovatnoće sekvenci, koje su od interesa. Matematički aparat vezan za modele jezika je uglavnom nezavisan od jezika. Međutim, kvalitet obučenih modela ne zavisi samo od algoritama obuke, već prvenstveno od količine i kvaliteta podataka koji su na raspolaganju za obuku. Za jezike sa kompleksnom morfologijom, kao što je srpski, tekstualni korpus za obuku modela mora biti daleko obimniji od korpusa koji bi se koristio kod nekog od jezika sa relativno jednostavnom morfologijom, poput engleskog. Ovo istraživanje obuhvata razvoj jezičkih modela za srpski jezik, počevši od prikupljanja i inicijalne obrade tekstualnih sadržaja, preko adaptacije algoritama i razvoja metoda za rešavanje problema nedovoljne količine podataka za obuku, pa do prilagođavanja i primene modela u različitim tehnologijama, kao što su sinteza govora na osnovu teksta, automatsko prepoznavanje govora, automatska detekcija i korekcija gramatičkih i semantičkih grešaka u tekstovima, a postavljaju se i osnove za primenu jezičkih modela u automatskoj klasifikaciji dokumenata i drugim tehnologijama. Jezgro razvoja jezičkih modela za srpski predstavlja definisanje morfoloških klasa reči na osnovu informacija koje su sadržane u morfološkom rečniku, koji je nastao kao rezultat jednog od ranijih istraživanja.*

Datum prihvatanja teme od strane NN veća: **DP**	10.5.2018.
Datum odbrane: **DO**	
Članovi komisije: (ime i prezime / titula / zvanje / naziv organizacije / status) **KO**	**predsednik:** dr Vlado Delić redovni profesor Fakultet tehničkih nauka, Novi Sad **član:** dr Dragana Bajić redovni profesor Fakultet tehničkih nauka, Novi Sad **član:** dr Snežana Gudurić redovni profesor Filozofski fakultet, Novi Sad **član:** dr Tatjana Grbić vanredni profesor Fakultet tehničkih nauka, Novi Sad **član:** dr Jelena Nikolić docent Elektronski fakultet, Niš **član, mentor:** dr Milan Sečujski vanredni profesor Fakultet tehničkih nauka, Novi Sad
	Potpis mentora:

UNIVERSITY OF NOVI SAD
FACULTY OF TECHNICAL SCIENCES

KEY WORDS DOCUMENTATION

Accession number: **ANO**	
Identification number: **INO**	
Document type: **DT**	Monograph documentation
Type of record: **TR**	Textual printed material
Contents code: **CC**	PhD thesis
Author: **AU**	Stevan Ostrogonac, M.Sc.
Mentor: **MN**	prof. Milan Sečujski, PhD
Title: **TI**	Models of the Serbian language and their application in speech and language technologies
Language of text: **LT**	Serbian
Language of abstract: **LA**	Serbian/English
Country of publication: **CP**	Serbia
Locality of publication: **LP**	Vojvodina
Publication year: **PY**	2018.
Publisher: **PU**	Faculty of Technical Sciences

Publication place: **PP**	Trg Dositeja Obradovića 6, Novi Sad
Physical description: **PD**	6 chapters / 98 pages / 0 citations / 15 figures / 10 graphs / 19 tables / 89 references / 5 appendices
Scientific field **SF**	Electrical and Computer Engineering
Scientific discipline **SD**	Telecommunications
Subject, Key words **SKW**	Natural language processing Computational linguistics
UC	
Holding data: **HD**	in the library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad
Note: **N**	
Abstract: **AB**	*A statistical language model, in theory, represents a probability distribution over sequences of words of a language. In practice, it is a tool for estimating probabilities of word sequences of interest. Mathematical basis related to language models is mostly language independent. However, the quality of trained models depends not only on training algorithms, but on the amount and quality of available training data as well. For languages with complex morphology, such as Serbian, textual corpora for training language models need to be significantly larger than the corpora needed for training language models for languages with relatively simple morphology, such as English. This research represents the entire process of developing language models for Serbian, starting with collecting and preprocessing of textual contents, extending to adaptation of algorithms and development of methods for addressing the problem of insufficient training data, and finally to adaptation and application of the models in different*

	technologies, such as text-to-speech synthesis, automatic speech recognition, automatic detection and correction of grammar and semantic errors in texts, and determining basics for the application of the models in automatic document classification and other tasks. The core of the development of language models for Serbian is defining morphologic classes of words, based on the information contained within the morphologic dictionary of Serbian, which was one of the results of a previous research.
Accepted on Scientific Board on: **AS**	5/10/2018
Defended: **DE**	
Thesis Defend Board: **DB**	**president:** Vlado Delić, PhD full professor Faculty of Technical Sciences, Novi Sad **member:** Snežana Gudurić, PhD full professor Faculty of Philosophy, Novi Sad **member:** Dragana Bajić, PhD full professor Faculty of Technical Sciences, Novi Sad **member:** Tatjana Grbić, PhD associate professor Faculty of Technical Sciences, Novi Sad **member:** Jelena Nikolić, PhD assistant professor Faculty of Electronic engineering, Niš **member, mentor:** Milan Sečujski, PhD associate professor Faculty of Technical Sciences, Novi Sad
	Mentor's sign:

SADRŽAJ

SPISAK SLIKA

SPISAK TABELA

SAŽETAK

Jezički modeli su sastavni deo mnoštva aplikacija baziranih na govornim i jezičkim tehnologijama. Zadatak ovih modela je određivanje verovatnoće datih sekvenci reči, najčešće radi određivanja najverovatnije sekvence iz skupa ponuđenih, kao što je, na primer, slučaj kod automatskog prepoznavanja govora, mašinskog prevođenja, ili automatske detekcije i korekcije grešaka u tekstualnim sadržajima. Za obuku jezičkih modela potrebni su obimni tekstualni korpusi. Prikupljanje kvalitetnih tekstualnih sadržaja je vremenski veoma zahtevan posao, te za mnoge jezike za sada ne postoje adekvatni resursi za obuku jezičkih modela. Problem nedovoljne količine podataka za obuku jezičkih modela je naročito izražen kod jezika sa relativno kompleksnom morfologijom. U ovu grupu jezika spada i srpski. Problem nedostatka podataka za obuku najčešće se tretira klasterizacijom reči na osnovu statističkih podataka dobijenih iz tekstualnog korpusa. Za određivanje verovatnoća određenih sekvenci reči, kod ovakvih modela jezika se koriste verovatnoće sekvenci klasa reči. Međutim, za jezike sa kompleksnom morfologijom, klasterizacija na osnovu statistika vezanih za najčešće kontekste u kojima se reči pojavljuju ne predstavlja adekvatno rešenje, s obzirom na velik broj reči koje se u korpusu pojavljuju veoma retko, u mnogim slučajevima i samo jednom. Stoga je potrebno adaptirati tehnike modelovanja kako bi se postigla dobra reprezentacija jezika. Istraživanje koje je opisano u ovoj disertaciji predstavlja razvoj jezičkih modela za srpski jezik, koji uključuje prikupljanje i obradu tekstualnih sadržaja, razvoj tehnika modelovanja kojima se uzimaju u obzir morfološke informacije, kao i adaptaciju i primenu jezičkih modela u govornim i jezičkim tehnologijama.

U uvodnom poglavlju objašnjen je pojam jezičkog modela, njegova uloga u govornim i jezičkim tehnologijama, kao i uobičajeni načini evaluacije kvaliteta modela. U drugom poglavlju prikazano je stanje u oblasti, čime su obuhvaćene i matematičke osnove dve najrasprostranjenije paradigme jezičkog modelovanja – statistički *N*-gram modeli i modeli bazirani na neuronskim mrežama. U trećem poglavlju opisan je proces prikupljanja i obrade tekstualnih korpusa za srpski jezik, sa osvrtom na ostale jezičke resurse, koji su korišćeni u okviru ovog istraživanja. U četvrtom poglavlju opisan je postupak grupisanja reči na osnovu morfoloških informacija, kao i modeli koji su zasnovani na ovakvim klasama. U petom poglavlju predstavljeni su rezultati istraživanja vezani za adaptaciju i primenu jezičkih modela za srpski jezik u govornim i jezičkim tehnologijama. U šestom poglavlju sumirani su rezultati ovog istraživanja i obrazložen je njihov značaj za dalji razvoj govornih tehnologija za srpski jezik.

ABSTRACT

Language models are important parts of a variety of applications that are based on speech and language technologies. The purpose of these models is determining probabilities of given word sequences, usually in order to select the most probable word sequence from a given set of word sequences, as is the case in automatic speech recognition, machine translation or in automatic detection and correction of errors in textual contents. Training language models requires large textual corpora. Collecting quality textual content is a very time-consuming task, which is the reason why, for many languages, there are no adequate resources for language modeling to date. The problem of insufficient training data is additionally emphasized for languages with relatively complex morphology. The Serbian language falls into this group. The problem of data sparsity is commonly treated by word clustering based on statistics derived from training corpora. With these models, estimation of word sequence probabilities is performed by using class sequence probabilities. However, for languages with complex morphology, word clustering based on statistics related to the most frequent contexts in which the words appear in the training corpora is not an adequate solution, since many words appear very rarely in the corpus, often only once. Therefore, in order to achieve a fair language representation, it is necessary to adapt modeling techniques. The research presented within this dissertation gives an overview of the development of language models for the Serbian language, which includes collecting and preprocessing of textual contents, development of modeling techniques that take into account morphologic information, as well as adaptation and application of language models in speech and language technologies.

Within the introductory chapter, the term language model is explained in more detail, as well as its role in speech and language technologies and common techniques for determining how well a model represents a language. Current state in the field is given in the second chapter, which includes mathematical basics related to the two mainstream language modeling paradigms – statistical *N*-gram models and models based on neural networks. In the third chapter, the process of collecting and preprocessing of textual contents in Serbian is described, along with other language resources that were used within this research. The fourth chapter gives a detailed description of word clustering based on morphologic information and a description of language models that are based on morphologic word classes. The fifth chapter presents the results of research related to the adaptation and application of language models for Serbian in speech and language technologies. The document is concluded with chapter six, in which the research results are summarized and their impact on further development of speech and language technologies for Serbian is explained.

ZAHVALNICA

Zahvaljujem se prof. dr Vladi Deliću, koji me je uveo u svet nauke i omogućio mi da se bavim poslom koji volim, na savetima i na pomoći koju mi je pružao tokom studija.

Na izvanredno zanimljivim predavanjima, koja su me, još tokom osnovnih studija, zainteresovala za govorne tehnologije, a kasnije na pouzdanoj podršci koju sam imao tokom istraživanja i pisanja naučnih publikacija, zahvaljujem se mentoru, prof. dr Milanu Sečujskom.

Veliku zahvalnost dugujem i kolegama – Robertu Maku, Dragiši Miškoviću, dr Branislavu Popoviću, dr Branku Lučiću, mr Nataši Vujnović-Sedlar, Edvinu Pakociju, Siniši Suziću, kao i mnogim drugima, na prijatnoj i prijateljskoj saradnji tokom istraživanja.

Takođe, želim da se zahvalim preduzeću „AlfaNum“ iz Novog Sada, na ustupljenoj programskoj podršci i resursima, koji su mi bili od velike pomoći u toku istraživanja.

Posebnu zahvalnost upućujem svojoj porodici, a naročito supruzi Anđi, na strpljenju, razumevanju, podršci i ljubavi koju mi pruža.

POGLAVLJE 1

UVOD

Modelovanje jezika predstavlja jedan od zadataka računarske nauke koja se naziva obrada prirodnog jezika (eng. *natural language processing – NLP*). Obrada prirodnog jezika počela je da se razvija sredinom prošlog veka, kada su kreirani prvi sistemi sa ciljem da učine komunikaciju između čoveka i mašine što sličnijom komunikaciji korišćenjem prirodnog govornog jezika (Weizenbaum, 1966). Prvi ovakvi sistemi su se oslanjali na skup manuelno implementiranih pravila i bili su veoma ograničenih mogućnosti. Vremenom, paralelno sa razvojem računarskih tehnologija, razvijani su i sistemi koji su se oslanjali na baze podataka (tekstualne korpuse), na osnovu kojih su, putem statističke analize, automatski ili poluautomatski izvođena pravila govorne komunikacije (Manning & Schütze, 1999). U tom periodu došlo je do značajnih pomaka u rešavanju mnogih problema koji pripadaju oblasti NLP. Definisane su nove tehnike, u kojima su, između ostalog, primenjivani matematički koncepti koji su već bili u upotrebi u drugim istraživačkim oblastima (klasifikaciona i regresiona stabla, skriveni Markovljevi modeli itd.), a pomoću kojih je obrada prirodnog jezika nastavila da se razvija na različitim nivoima – fonološkom, morfološkom, sintaksnom, semantičkom, kao i na nivou diskursa. Poslednjih godina, velika pažnja posvećena je neuronskim mrežama, naročito strukturama kao što su tzv. duboke neuronske mreže (eng. *deep neural networks – DNN*), koje su se pokazale pogodnim za modelovanje komplikovanih međuzavisnosti jezičkih segmenata, naročito onih hijerarhijske prirode (Mikolov, 2012).

Jezički model (eng. *language model – LM*), u teoriji, predstavlja raspodelu verovatnoća nad skupom sekvenci reči nekog jezika. Ovo važi kada se govori o statističkim modelima. U praksi se, naravno, vrše samo estimacije verovatnoća pojedinih sekvenci reči, a najčešći zadatak jezičkog modela je zapravo određivanje najverovatnije sekvence reči iz skupa ponuđenih sekvenci (hipoteza). Modeli jezika se u različitim oblicima koriste u govornim i jezičkim tehnologijama. Dva najrasprostranjenija tipa su jezički modeli bazirani na statistikama tzv. *N*-grama reči i modeli bazirani na neuronskim mrežama. Matematički aparat vezan za modele jezika je uglavnom nezavisan od jezika. Međutim, kvalitet obučenih modela nije zavisan samo od algoritama obuke, već prvenstveno od količine i kvaliteta podataka koji su na raspolaganju za obuku. Za jezike sa kompleksnom morfologijom, kao što je srpski, tekstualni korpus za obuku modela mora biti daleko obimniji od korpusa koji bi se koristio kod nekog od jezika sa relativno jednostavnom morfologijom, poput npr. engleskog. Prikupljanje i obrada podataka

za obuku vremenski je veoma zahtevan posao, te za mnoge jezike, pa ni za srpski, ne postoje adekvatni resursi za izgradnju kvalitetnih jezičkih modela, makar kada su u pitanju primene za koje se zahteva rad sa velikim rečnicima.

Kada je u pitanju određivanje kvaliteta jezičkog modela, razlikuju se dve grupe mera – intrinsične (nezavisne od primene modela) i ekstrinsične (vezane za konkretnu primenu) (Chen et al, 1998). Najčešće korišćena intrinsična mera kvaliteta je tzv. perpleksnost (eng. *perplexity* – *ppl*). U kontekstu jezičkog modelovanja, perpleksnost predstavlja recipročnu vrednost verovatnoće sekvence reči, koja predstavlja skup podataka za testiranje, normalizovanu dužinom te sekvence. Naime, ako se govornik posmatra kao diskretan izvor informacija, koji generiše sekvencu reči $w_1, w_2, \dots, w_m$ (koje pripadaju rečniku W), pri čemu verovatnoća reči w_i zavisi od niza reči koje joj prethode $w_1, w_2, \dots, w_{i-1}$, entropija ovakvog izvora, izražena na nivou reči, predstavlja prosečnu količinu informacije koju donosi nova reč:

$$H = -\lim_{m\to\infty} \frac{1}{m} \sum_{w_1,w_2,\dots,w_m} (P(w_1, w_2, \dots, w_m) \log_2 P(w_1, w_2, \dots, w_m)). \quad (1.1)$$

Sumira se po svim mogućim sekvencama reči, ali pod pretpostavkom ergodičnosti, izraz (1.1) može se napisati u obliku:

$$H = -\lim_{m\to\infty} \frac{1}{m} \log_2 P(w_1, w_2, \dots, w_m). \quad (1.2)$$

Ova pretpostavka proizilazi iz činjenice da se jezikom uspešno možemo služiti, iako nismo prethodno imali prilike da čujemo sve reči koje su ikada bile napisane ili izgovorene. Takođe, značenje pojedinih reči čovek može utvrditi na osnovu malog segmenta prethodnog teksta. Dalje, za dovoljno veliko m, entropija se može estimirati na sledeći način:

$$\hat{H} = -\frac{1}{m} \log_2 P(w_1, w_2, \dots, w_m). \quad (1.3)$$

Perpleksnost se, kao mera kvaliteta jezičkog modela, definiše kao:

$$ppl = 2^{\hat{H}} = \hat{P}(w_1, w_2, \dots, w_m)^{-\frac{1}{m}}. \quad (1.4)$$

U izrazu (1.4), $\hat{P}$ predstavlja estimiranu verovatnoću koja je određenoj sekvenci reči dodeljena od strane jezičkog modela. Za statističke N-gram modele jezika, proračun perpleksnosti se, primenom lančanog pravila, svodi na sledeći postupak:

$$ppl = \sqrt[m]{\frac{1}{\prod_{i=1}^{m} P(w_i | w_{i-N+1}, \dots, w_{i-1})}}. \quad (1.5)$$

U izrazu (1.5), treba napomenuti da N predstavlja red modela (za trigram model), $N = 3$, što znači da je relevantan kontekst sačinjen od prethodnih $N - 1$ reči. U literaturi se može sresti i definicija reda modela kao dužine konteksta, ali će u okviru ovog teksta biti korišćena gore pomenuta konvencija. Iako je perpleksnost često korišćena mera za procenu kvaliteta modelovanja jezika, pokazuje se da u određenim aplikacijama perpleksnost nije značajno korelisana sa stvarnim performansama modela jezika. Od ekstrinsičnih mera kvaliteta najčešće se koristi stopa pogrešno prepoznatih reči u sistemima za automatsko prepoznavanje govora (eng. *word error rate – WER*). Ova mera računa se prema sledećem izrazu:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}, \tag{1.6}$$

pri čemu u izrazu (1.6) S predstavlja broj zamena (izgovorena reč prepoznata je kao neka druga), D predstavlja broj brisanja (izgovorena reč nije detektovana), I predstavlja broj umetanja (detektovana je reč, iako nije izgovorena), C je broj korektno identifikovanih reči, a N je broj reči u referentnom tekstu. Iako se *WER* (tačnije, smanjenje *WER* uvođenjem jezičkog modela u proces prepoznavanja govora) smatra objektivnom merom kvaliteta jezičkog modela, činjenica je da se u *WER* ogleda komplementarnost akustičkog i jezičkog modela, kao i reprezentativnost uzorka koji se koristi za testiranje, zbog čega je često pogodno koristiti modele jezika koji su prilagođeni konkretnoj aplikaciji (Mikolov et al, 2011a). U takvim situacijama rešavanje problema nedovoljne količine podataka za obuku modela postaje naročito važno.

Istraživanje koje je predmet ove disertacije obuhvata razvoj jezičkih modela za srpski jezik, počevši od prikupljanja i inicijalne obrade tekstualnih sadržaja, preko adaptacije algoritama i razvoja metoda za rešavanje problema nedostajućih podataka za obuku, pa do prilagođavanja i primene modela u različitim tehnologijama, koje uključuju sintezu govora na osnovu teksta, automatsko prepoznavanje govora, i korekciju gramatičkih grešaka u tekstovima, a postavljaju se i osnove za primenu jezičkih modela u automatskoj klasifikaciji dokumenata i drugim tehnologijama. Jezgro razvoja jezičkih modela za srpski predstavlja definisanje morfoloških klasa reči na osnovu informacija koje su sadržane u postojećem morfološkom rečniku. Ove klase, u slučajevima nedostajućih podataka za obuku, zamenjuju konkretne reči pri obuci jezičkih modela. U okviru istraživanja koje je predmet ove disertacije, grupisanje reči na osnovu morfoloških informacija poređeno je sa standardnim načinom grupisanja reči za potrebe modelovanja jezika, koji se oslanja na statistike tekstualnih korpusa. Grupisanje na osnovu morfoloških rezultata pokazalo se kao adekvatnije rešenje. Modeli koji su obučavani na korpusima morfoloških klasa mogu se, za određene primene, kombinovati sa modelima koji su obučavani na korpusima reči. Reči se takođe mogu zameniti odgovarajućim osnovnim oblicima reči – lemama,

što predstavlja međustepen u odnosu na klasifikaciju reči na osnovu morfoloških informacija.

U poglavlju 2 predstavljene su dve najrasprostranjenije paradigme jezičkog modelovanja – statistički *N*-gram modeli i modeli bazirani na neuronskim mrežama. Prikazane su najčešće korišćene tehnike za obuku modela, njihovo kombinovanje, tehnike kojima se modeli mogu prilagoditi određenim primenama, a predstavljeni su i najpoznatiji alati koji se koriste u navedene svrhe. U poglavlju 3 predstavljeni su resursi koji su potrebni za kreiranje kvalitetnih jezičkih modela za srpski jezik, kako oni koji su razvijeni ranije, tako i oni koji su nastali kao deo istraživanja koje je predmet ove disertacije. Detaljno je opisan proces obrade tekstualnih sadržaja na srpskom jeziku sa svim pratećim problemima, kao i alati koji su nastali tokom tog procesa. U poglavlju 4 opisan je način na koji su definisane klase reči na osnovu morfoloških informacija, objašnjena je struktura jezičkih modela koji se baziraju na ovim klasama, kao i alati koji su razvijeni za potrebe obuke i primene ovakvih modela. Poglavlje 5 predstavlja pregled istraživanja vezanih za modelovanje jezika i njihovu primenu u govornim i jezičkim tehnologijama za srpski jezik. U zaključku su sumirani rezultati i doprinosi ove disertacije, a naznačeni su i pravci daljeg istraživanja.

POGLAVLJE 2

JEZIČKI MODELI – STANJE U OBLASTI

Od najrasprostranjenijih paradigmi jezičkog modelovanja, u proteklim decenijama najviše tehnika za poboljšanje osnovnih struktura modela razvijeno je u okviru dve – statističkog modelovanja na osnovu *N*-grama i modelovanja zasnovanog na neuronskim mrežama. Jezički modeli zasnovani na neuronskim mrežama (eng. *neural network language model* – *NNLM*) privukli su pažnju istraživača tek u proteklih desetak godina. Razlozi za ovo leže u ranijim debatama o mogućnostima i ograničenjima neuronskih mreža, ali pre svega u činjenici da je za obuku neuronskih mreža potrebna veoma velika količina podataka, iako je i za *N*-gram modele neophodno posedovati obimne tekstualne korpuse. Pored toga, za obuku modela baziranih na neuronskim mrežama potrebni su računarski resursi kakvi su postali dostupni tek u vrlo bliskoj prošlosti. Uprkos porastu procesorske moći i memorijskih kapaciteta savremenih računara, mnoge aplikacije kreiraju se za uređaje koji raspolažu daleko skromnijim resursima u odnosu na standardne računare. Iz tog razloga, između ostalog, *N*-gram modeli se i dalje često koriste. Takođe, utvrđeno je da se u pojedinim situacijama *N*-gram modeli pokazuju uspešnijim od modela baziranih na neuronskim mrežama, o čemu će biti reči u odeljku 2.3.

2.1 Statistički *N*-gram modeli jezika

Modeli jezika koriste se u različitim tehnologijama i najčešće im je uloga određivanje najverovatnije sekvence reči iz skupa ponuđenih. U automatskom prepoznavanju govora to može biti skup najverovatnijih hipoteza, koji se dobija na izlazu akustičkih modela. U mašinskom prevođenju takođe je potrebno odrediti koja je najadekvatnija sekvenca reči na ciljnom jeziku koja odgovara sekvenci reči na polaznom jeziku. Kod detekcije i korekcije grešaka u tekstualnim sadržajima, situacija je nešto drugačija. Naime, detekcija grešaka u tekstovima predstavlja poseban izazov jer je potrebno odrediti na kojim mestima u tekstu je verovatnoća određene reči u datom kontekstu dovoljno mala da bi predstavljala indikator greške. Potrebno je, dakle, obučiti sistem u smislu finog podešavanja parametara, radi što tačnije detekcije grešaka. Tek kada je greška detektovana, moguće je generisati skup hipoteza, odnosno reči ili sekvenci reči kojima bi se trebala izvršiti supstitucija.

Bez obzira na to kakva je konkretna primena u pitanju, zadatak jezičkog modela jeste predviđanje reči w_n ako je poznat kontekst (istorija), odnosno prethodnih $N-1$ reči. Drugim rečima, izlaz jezičkog modela treba da bude uslovna verovatnoća $P(w_n|w_1,\dots,w_{n-1})$. U praksi, broj različitih istorija je isuviše velik da bi se ovakve verovatnoće mogle izračunati. Stoga se istorije moraju na neki način klasterizovati kako bi se omogućila estimacija pomenute verovatnoće. Jedan od načina da se ovo izvede je pomoću Markovljeve pretpostavke, odnosno, pretpostavke da je za predikciju određene reči dovoljno poznavati lokalni kontekst, odnosno prethodnih nekoliko reči. U praksi se najčešće koriste istorije dužine 1 (bigram modeli), 2 (trigram modeli) i 3 (kvadrigram modeli). Za, na primer, trigram modele, zadatak je estimacija verovatnoća $P(w_n|w_{n-2},w_{n-1})$. Pored ovakvog načina klasterizacije istorija, ređe se koriste modeli jezika zasnovani na *N*-gramima slova ili grafema uopšte, *N*-gramima koji se formiraju od korena pojedinih reči, kao i različite tehnike zasnovane na izdvajanju dodatnih informacija iz teksta (Manning & Schütze, 1999).

2.1.1 Određivanje verovatnoća *N*-grama

Obuka statističkih *N*-gram modela jezika predstavlja proces određivanja uslovnih verovatnoća $P(w_n|w_{n-N+1},\dots,w_{n-1})$, gde je N red modela, definisan po konvenciji koja je uvedena u uvodnom poglavlju. Radi jednostavnosti, u daljem tekstu će se pod verovatnoćama *N*-grama podrazumevati upravo ove uslovne verovatnoće, iako se u literaturi češće pod verovatnoćom *N*-grama podrazumevaju verovatnoće $P(w_{n-N+1},\dots,w_n)$.

Na osnovu verovatnoća *N*-grama, moguće je (primenom lančanog pravila uz već pomenutu Markovljevu pretpostavku) odrediti verovatnoću bilo koje sekvence reči *W* dužine *K*, pomoću izraza:

$$P(W) = P(w_1,w_2,\dots,w_K) = P(w_1)P(w_2|w_1)\dots P(w_{N-1}|w_1,\dots,w_{N-2})\prod_{i=N}^{K} P(w_i|w_{i-N+1},\dots,w_{i-1}). \quad (2.1)$$

Naravno, određena reč zavisi u izvesnoj meri i od reči koje nisu obuhvaćene kontekstom dužine 2 ili 3, ali se modelovanje jezika na ovaj način ipak pokazalo dovoljno dobrim za većinu primena. U odeljku 2.4 biće više reči o različitim vrstama *N*-gram modela i načinima na koje se dodatno poboljšavaju performanse osnovnih modela.

Verovatnoće *N*-grama određuju se pomoću velikog tekstualnog korpusa. Na osnovu definicije uslovne verovatnoće, dobija se:

$$P(w_n|w_1, \dots, w_{n-1}) = \frac{C(w_1,..,w_n)}{C(w_1,...,w_{n-1})}, \quad (2.2)$$

pri čemu u izrazu (2.2) C predstavlja broj pojavljivanja određene sekvence reči. Za sve N-grame koji se završavaju rečju w_n, pri čemu se ta reč ne pojavljuje u korpusu za obuku nakon sekvence $w_1, \dots, w_{n-1}$ dodeljena verovatnoća bila bi 0. Zbog toga bi, s obzirom na način računanja verovatnoća dužih sekvenci (množenjem verovatnoća N-grama), većini dužih sekvenci bila dodeljena verovatnoća 0. Pored toga, ako se sekvenca $w_1, \dots, w_{n-1}$ pojavljuje tačno jednom, ili više puta, ali nakon nje uvek sledi ista reč w_n, ovom N-gramu bila bi dodeljena verovatnoća 1. Problem nedovoljne količine podataka za obuku karakterističan je praktično za sve zadatke u okviru obrade prirodnog jezika, a radi delimičnog rešavanja ovog problema koriste se metode tzv. ublažavanja (eng. *smoothing*) raspodela verovatnoća. Neke od njih, koje se najčešće koriste pri modelovanju jezika pomoću N-grama, biće izložene u narednom odeljku.

2.1.2 TEHNIKE UBLAŽAVANJA RASPODELE VEROVATNOĆA

Postoji mnoštvo tehnika ublažavanja raspodele verovatnoća, pri čemu se ove tehnike razlikuju po kompleksnosti, a u skladu sa tim i po doprinosu kvalitetu jezičkih modela za koje se koriste (Chen & Goodman, 1999).

Ublažavanje dodavanjem

Ublažavanje raspodele verovatnoće dodavanjem određenog broja ukupnom broju pojavljivanja svakog od N-grama predstavlja najjednostavniju tehniku, ali i daje znatno lošije rezultate od drugih tehnika. Računanje verovatnoća N-grama se izvodi prema sledećem izrazu:

$$P(w_i|w_{i-N+1}, \dots, w_{i-1}) = \frac{\delta + C(w_{i-N+1},...,w_i)}{\delta|V| + C(w_{i-N+1},...,w_{i-1})}, \quad (2.3)$$

kada navedene sekvence postoje, pri čemu u izrazu (2.3) δ predstavlja broj koji se dodaje stvarnom broju pojavljivanja N-grama u korpusu za obuku modela, dok je $|V|$ veličina rečnika koji se razmatra. Za δ se u praksi često koristi vrednost 1 (eng. *Add-1 Smoothing*), ali ako postoji reprezentativan skup podataka za optimizaciju, onda se ova vrednost može prilagoditi tom skupu (eng. *Add-δ Smoothing*). Na ovaj način izbegava se dodeljivanje verovatnoće 0 N-gramima koji se ne pojavljuju u korpusu za obuku. S druge strane, model jezika koji se obučava na ovaj način dodeljuje istu verovatnoću svim ovakvim N-gramima. Ovaj problem se može

delimično rešiti interpolacijom sa unigram modelom (Klakow, 1998), čime se uzima u obzir broj pojavljivanja reči w_i.

Good-Turing ublažavanje

Ideja *Good-Turing* ublažavanja jeste da se za *N*-grame koji su se u korpusu za obuku pojavili *r* puta odvoji deo mase verovatnoće od *N*-grama koji su se u istom korpusu pojavili $r+1$ put, kako bi se odredili očekivani brojevi pojavljivanja *N*-grama – r^* (Good, 1953).

Počevši od toga da se svaki *N*-gram α_i u korpusu pojavljuje sa verovatnoćom p_i (koja nije poznata) i pretpostavke da su pojavljivanja određenog *N*-grama u korpusu međusobno nezavisna, kao i pretpostavke da je broj pojavljivanja α_i, $c(\alpha_i)$, u skladu sa binomnom raspodelom, važi:

$$p(c(\alpha_i)=r) = b(r;N,p_i) = \binom{N}{r} p_i^r (1-p_i)^{N-r}. \quad (2.4)$$

U izrazu (2.4) N predstavlja ukupan broj različitih *N*-grama, koji se pojavljuju u korpusu za obuku. Pošto je cilj da se izračuna očekivani broj pojavljivanja (r^*, odnosno $E\big(c^*(\alpha_i)\big)$), koristi se sledeći izraz:

$$E\big(c^*(\alpha_i)\big) = Np_i. \quad (2.5)$$

Ako se u korpusu pojavljuje s različitih *N*-grama $(\alpha_1,\dots,\alpha_s)$, sa verovatnoćama $p_1,\dots,p_s$, očekivani broj *N*-grama koji se pojavljuju r puta u korpusu možemo izraziti na sledeći način:

$$E_N(N_r) = \sum_{i=1}^{s} p\big(c(\alpha_i=r)\big) = \sum_{i=0}^{s}\binom{N}{r} p_i^r (1-p_i)^{N-r}. \quad (2.6)$$

Za male vrednosti *r*, može se pretpostaviti da je očekivani broj *N*-grama koji se pojavljuju *r* puta približno jednak broju *N*-grama koji se zaista pojavljuju *r* puta u tekstualnom korpusu. Verovatnoća da neki *N*-gram α spada u grupu α_i, ako je broj pojavljivanja α_i jednak *r*, je:

$$P(\alpha=\alpha_i | c(\alpha_i)=r) = \frac{P(c(\alpha_i)=r)}{\sum_{j=1}^{s} P(c(\alpha_j)=r)}. \quad (2.7)$$

Očekivani broj pojavljivanja ovog *N*-grama je:

$$E(c^*(\alpha_i)|c(\alpha_i)=r) = \sum_{i=1}^{s} Np_i p(\alpha=\alpha_i|c(\alpha_i)=r). \quad (2.8)$$

Iz (2.7) i (2.8) proizilazi da je:

$$E(c^*(\alpha_i)|c(\alpha_i)=r)=\sum_{i=1}^{s}Np_i\frac{P(c(\alpha_i)=r)}{\sum_{j=1}^{s}P(c(\alpha_j)=r)}=\frac{\sum_{i=1}^{s}Np_ip(c(\alpha_i)=r)}{\sum_{j=1}^{s}p(c(\alpha_j)=r)}. \quad (2.9)$$

Brojilac u izrazu (2.9) predstavlja zapravo očekivani broj *N*-grama koji se pojavljuju *r* puta. Što se tiče imenioca, važi sledeće:

$$\begin{aligned}\sum_{i=1}^{s}Np_ip(c(\alpha_i)=r)&=\sum_{i=1}^{s}Np_i\binom{N}{r}p_i^r(1-p_i)^{N-r}\\&=N\frac{N!}{N-r!\,r!}p_i^{r+1}(1-p_i)^{N-r}\\&=N\frac{r+1}{N+1}\frac{N+1!}{N-r!\,r+1!}p_i^{r+1}(1-p_i)^{N-r}\\&=(r+1)\frac{N}{N+1}E_{N+1}(N_{r+1})\\&\cong(r+1)E_{N+1}(N_{r+1}). \quad (2.10)\end{aligned}$$

Konačno, dobija se da je:

$$r^*=E(c^*(\alpha_i)|c(\alpha_i)=r)=\frac{(r+1)E_{N+1}(N_{r+1})}{E_N(N_r)}\cong(r+1)\frac{N_{r+1}}{N_r}. \quad (2.11)$$

Na opisani način, svim *N*-gramima koji se nisu pojavili u korpusu za obuku dodeljuje se jednaka verovatnoća. Takođe, za veće *r* može se desiti da N_{r+1} bude 0, pa da modifikovan broj pojavljivanja ovakvih *N*-grama – r^* bude takođe 0. Iz oba razloga se *Good-Turing* metoda ublažavanja u praksi najčešće koristi kao osnova za neku od naprednijih tehnika.

Jelinek-Mercer ublažavanje

Jedan od načina da se napravi distinkcija među *N*-gramima koji se ne pojavljuju u korpusu za obuku jeste da se na neki način iskoriste informacije koje se mogu dobiti pomoću modela nižeg reda. *Jelinek-Mercer* ublažavanje upravo predstavlja linearnu interpolaciju osnovnog modela sa modelom nižeg reda (Jelinek & Mercer, 1985). Rekurzivna formulacija ovakvog pristupa glasi: Model reda *N* dobija se pomoću linearne interpolacije osnovnog modela reda *N* i modela reda $N-1$ na koji je već primenjeno *Jelinek-Mercer* ublažavanje. Ovo se može ilustrovati pomoću sledećeg izraza:

$$p_{JM}\left(w_i|w_{i-N+1}^{i-1}\right)$$
$$= \lambda_{w_{i-N+1}^{i-1}} p\left(w_i|w_{i-N+1}^{i-1}\right) + \left(1 - \lambda_{w_{i-N+1}^{i-1}}\right) p_{JM}\left(w_i|w_{i-N+2}^{i-1}\right), \quad (2.12)$$

pri čemu p_{JM} predstavljaju verovatnoće koje daju modeli na koje je primenjeno *Jelinek-Mercer* ublažavanje, dok p predstavljaju verovatnoće koje se dobijaju na osnvou modela na koji nije primenjeno ublažavanje, a λ predstavljaju težinske faktore. Oznake poput w_{i-N+1}^{i-1} predstavljaju samo skraćeni zapis sekvence reči (za dati primer, u pitanju je skraćena oznaka za sekvencu $w_{i-N+1}, \ldots, w_{i-1}$). Za bigram modele radi se interpolacija sa unigram modelom na koji nije primenjeno ublažavanje, ili unigram modelom za koji je ublažavanje urađeno nekom od postojećih tehnika.

Koeficijenti λ mogu se zadati manuelno, ali se u praksi najčešće za fino podešavanje ovih parametara koristi ili odvojen skup podataka za obuku (eng. *held-out data*), uz primenu algoritma maksimizacije očekivanja (eng. *Expectation Maximization – EM*), ili se vrši unakrsna validacija primenom tzv. interpolacije sa brisanjem (eng. *deleted interpolation*). Generalno, potrebno je učiniti da konteksti sa velikom frekvencijom pojavljivanja u korpusu za obuku dobiju veće vrednosti λ koeficijenata.

Katz ublažavanje, *back-off* model

Tehnika ublažavanja koju je predložio *Katz* (Katz, 1987) zasniva se na *Good-Turing* ublažavanju. Naime, svi brojevi pojavljivanja *N*-grama prvo se modifikuju u skladu sa *Good-Turing* postupkom. Na ovaj način realocirana masa verovatnoće raspoređuje se *N*-gramima koji se nisu pojavili u korpusu za obuku, ali u skladu sa raspodelom nižeg reda $(N-1)$. Formulacija *Katz* modela je rekurzivnog karaktera, tako da se radi ilustracije postupka može posmatrati samo bigram model.

Najpre je potrebno odrediti modifikovane vrednosti broja pojavljivanja *N*-grama, c_{Katz}:

$$c_{Katz}\left(w_{i-1}^{i}\right) = \begin{cases} d_r r & ako\ je\ r > 0 \\ \alpha(w_{i-1}) p(w_i) & ako\ je\ r = 0 \end{cases}, \quad (2.13)$$

gde je r stvarni broj pojavljivanja *N*-grama, a d_r koeficijent kojim je taj broj modifikovan u okviru *Katz* modela . Vrednosti $\alpha(w_{i-1})$ se biraju tako da:

$$\sum_{w_i} c_{Katz}\left(w_{i-1}^{i}\right) = \sum_{w_i} c\left(w_{i-1}^{i}\right), \quad (2.14)$$

na osnovu čega se dolazi do:

$$\alpha(w_{i-1}) = \frac{1-\sum_{w_i:c(w_{i-1}^{i})>0} p_{Katz}(w_i|w_{i-1})}{1-\sum_{w_i:c(w_{i-1}^{i})>0} p(w_i)}, \tag{2.15}$$

gde su sa p_{Katz} označene verovatnoće dobijene pomoću *Katz* modela, a sa p verovatnoće dobijene pomoću modela modela na koji nije primenjeno ublažavanje.

Vrednosti $\alpha(w_{i-1})$ iz izraza (2.15) se nazivaju *back-off* koeficijenti, s obzirom na to da se koriste pri prelasku na model nižeg reda. Verovatnoća *N*-grama se pomoću modifikovanih brojeva pojavljivanja *N*-grama može dobiti na sledeći način:

$$p_{Katz}(w_i|w_{i-1}) = \frac{c_{Katz}(w_{i-1}^{i})}{\sum_{wi} c_{Katz}(w_{i-1}^{i})}. \tag{2.16}$$

Što se tiče d_r, za veće brojeve pojavljivanja *N*-grama (npr. 5 ili više) najčešće se pretpostavlja da je broj pojavljivanja reprezentativan za konkretnu sekvencu reči, te se d_r postavlja na vrednost 1. Ako je *r* manje od određene granice, cilj je da umanjenje bude proporcionalno onom koje se dobija *Good-Turing* metodom:

$$1 - d_r = \mu\left(1 - \frac{r^*}{r}\right). \tag{2.17}$$

Drugim rečima, potrebno je da ukupna „masa“ broja pojavljivanja *N*-grama koja se realocira bude jednaka ukupnoj masi koja bi *Good-Turing* metodom bila dodeljena *N*-gramima koji se nisu pojavili u korpusu za obuku:

$$\sum_{r=1}^{k} N_r(1 - d_r)r = N_1, \tag{2.18}$$

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)N_{k+1}}{N_1}}{1-\frac{(k+1)N_{k+1}}{N_1}}. \tag{2.19}$$

Na ovaj način dobijaju se tzv. *Katz back-off* modeli. Za razliku od *Jelinek-Mercer* metode, ovde se, u opštem slučaju, ne vrši interpolacija modela različitih redova, već se koriste informacije modela nižih redova samo za *N*-grame koji se ne pojavljuju u korpusu za obuku.

Witten-Bell ublažavanje

Ova tehnika predstavlja poseban slučaj *Jelinek-Mercer* tehnike:

$$p_{WB}(w_i|w_{i-N+1}^{i-1}) = \lambda_{w_{i-N+1}^{i-1}} p(w_i|w_{i-N+1}^{i-1}) + (1-\lambda_{w_{i-N+1}^{i-1}}) p_{WB}(w_i|w_{i-N+2}^{i-1}), \quad (2.20)$$

pri čemu su sa p_{WB} označene verovatnoće *N*-grama koje se dobijaju na osnovu *Witten-Bell* modela, a sa *p* verovatnoće modela na koji nije primenjeno ublažavanje. Iako je reč o klasičnoj interpolaciji osnovnog modela i modela nižeg reda, polazi se od istog pristupa kao kod *Katz* modela, odnosno od ideje da se osnovni model koristi za *N*-grame koji se pojavljuju u korpusu za obuku, dok se za ostale koristi model nižeg reda. U skladu sa time, λ koeficijenti se interpretiraju kao verovatnoće da će u određenim situacijama biti korišćen osnovni model. Dakle, $1-\lambda$ predstavljaju verovatnoće da se reči koje se u korpusu za obuku nisu pojavile u određenom kontekstu pojave u tom kontekstu u skupu podataka za testiranje. Ove verovatnoće se estimiraju pomoću broja različitih koje se u datom kontekstu $(w_{i-N+1}, \dots, w_{i-1})$ pojavljuju u korpusu za obuku. U narednim izrazima, za broj različitih reči koje se pojavljuju u određenom kontekstu biće korišćen simbol „•“. Oznaka N_{1+} u narednim izrazima označava da se ono na šta se ta oznaka odnosi u tekstualnom korpusu pojavljuje jednom ili više puta.

$$N_{1+}(w_{i-N+1}^{i-1}\,\bullet) = \left|\{w_i : c(w_{i-N+1}^{i-1} w_i) > 0\}\right| \quad (2.21)$$

Potrebno je, dakle, postaviti koeficijente tako da važi:

$$1-\lambda_{w_{i-N+1}^{i-1}} = \frac{N_{1+}(w_{i-N+1}^{i-1}\bullet)}{N_{1+}(w_{i-N+1}^{i-1}\bullet)+\sum_{w_i} w_{i-N+1}^{i}}. \quad (2.23)$$

Ublažavanje apsolutnim umanjenjem

Ova metoda, kao i *Witten-Bell*, slična je *Jelinek-Mercer* ublažavanju, ali se pri interpolaciji modela ne vrši množenje verovatnoće modela višeg reda koeficijentom λ. Umesto toga, vrši se oduzimanje neke utvrđene vrednosti $\delta \in [0,1]$ od svih nenultih brojeva pojavljivanja. Izraz za računanje verovatnoće *N*-grama na osnovu modela koji se dobija ublažavanjem apsolutnim umanjenjem, p_{abs}, definisan je na sledeći način:

$$p_{abs}(w_i|w_{i-N+1}^{i-1}) = \frac{max\{c(w_{i-N+1}^{i})-\delta,0\}}{\sum_{w_i} c(w_{i-N+1}^{i})} + (1-\lambda_{w_{i-N+1}^{i-1}}) p_{abs}(w_i|w_{i-N+2}^{i-1}). \quad (2.24)$$

Kako bi se osiguralo da zbir verovatnoća bude 1, potrebno je izvršiti skaliranje na sledeći način:

$$1 - \lambda_{w_{i-N+1}^{i-1}} = \frac{\delta}{\sum_{w_i} c\left(w_{i-N+1}^{i}\right)} N_{1+}\left(w_{i-N+1}^{i-1} \bullet\right). \quad (2.25)$$

Važno je napomenuti da se parametar δ u praksi određuje pomoću skupa podataka koji se odvaja posebno u tu svrhu.

Kneser-Ney ublažavanje

Imajući u vidu činjenicu da je kod ublažavanja apsolutnim umanjenjem model nižeg reda najznačajniji u situacijama kada je broj pojavljivanja datog *N*-grama veoma mali ili 0, *Kneser-Ney* tehnika ublažavanja predstavlja adaptaciju apsolutnog ublažavanja u smislu prilagođavanja modela nižeg reda situacijama u kojima je on naročito od značaja. Pogodan primer za ilustraciju potrebe za ovom adaptacijom su sekvence *Stari Žednik* i *Novi Žednik*. Naime, može se dogoditi da se pomenute sekvence često pojavljuju u korpusu za obuku. Zbog toga, unigramu *Žednik* će na osnovu ublažavanja apsolutnim umanjenjem biti dodeljena velika verovatnoća. Međutim, s obzirom na to da se ovaj unigram pojavljuje gotovo uvek nakon reči *Stari* ili *Novi*, ukoliko bi se koristio u nekom novom kontekstu u okviru skupa za evaluaciju, verovatnoća koja mu je dodeljena bila bi neopravdano velika. Stoga je, kada se procenjuje verovatnoća neke reči, potrebno voditi računa i o broju različitih reči koje joj mogu prethoditi.

Ako se posmatraju bigram model i odgovarajući model nižeg reda – unigram, broj pojavljivanja svakog od unigrama može se izraziti pomoću broja različitih reči nakon kojih se on pojavljuje u korpusu za obuku:

$$N_{1+}(\bullet\, w_i) = |\{w_{i-1} : c(w_{i-1}w_i) > 0\}|. \quad (2.26)$$

Ukupan broj bigrama koji se pojavljuju u korpusu može se, koristeći do sada uvedenu notaciju, izraziti kao:

$$N_{1+}(\bullet\bullet) = \sum_{w_i} N_{1+}(\bullet\, w_i). \quad (2.27)$$

Pomoću izraza (2.26) i (2.27), dolazi se do izraza za verovatnoće pojedinih unigrama:

$$p_{KN}(w_i) = \frac{N_{1+}(\bullet w_i)}{N_{1+}(\bullet\bullet)}. \quad (2.28)$$

Modifikovani izraz za ublažavanje apsolutnim umanjenjem (p_{KN}) postaje:

$$p_{KN}\left(w_i | w_{i-N+1}^{i-1}\right) = \frac{max\left\{c\left(w_{i-N+1}^{i}\right) - \delta, 0\right\}}{\sum_{w_i} c\left(w_{i-N+1}^{i}\right)}$$

$$+\frac{\delta}{\sum_{w_i} c\left(w_{i-N+1}^{i}\right)} N_{1+}(w_{i-N+1}^{i-1}\bullet) p_{KN}\left(w_i | w_{i-N+2}^{i-1}\right). \qquad (2.29)$$

2.1.3 *SRILM TOOLKIT* – ALAT ZA OBUKU *N*-GRAM MODELA JEZIKA

Najpoznatiji alat koji se koristi za istraživanja vezana za *N*-gram modele jezika naziva se *SRILM toolkit* (*Stanford Research Institute Language Modeling toolkit*) i razvija se od 1995. godine (Stolcke, 2002). Ovaj alat sastoji se od mnoštva biblioteka *C++* kodova, koje su javno dostupne. Na osnovu ovih biblioteka, razvijena je grupa izvršnih programa, koji služe za obuku i testiranje različitih tipova jezičkih modela, njihovu interpolaciju i modifikaciju na razne načine. Takođe, javno su dostupni i jednostavni primeri korišćenja ovog alata u pomenute svrhe. Primer jednog *back-off* modela jezika koji se može napraviti uz pomoć *SRILM*-a prikazan je u prilogu 1. U nastavku će biti ukratko opisane funkcionalnosti pojedinih izvršnih programa *SRILM*-a.

ngram-count

Ovaj izvršni program predstavlja alat za obuku modela, odnosno izdvajanje statističkih informacija iz tekstualnog korpusa, primenu neke od tehnika ublažavanja o kojima je bilo reči u odeljku 2.1.2, a služi i za smanjenje veličine već obučenog modela ukoliko za tim postoji potreba. Pri tome, implementirane su brojne druge opcije kojima se detalji obuke mogu varirati sa ciljem da se dobije najpogodniji model za neku konkretnu primenu. Jedna on važnih opcija je i mogućnost definisanja rečnika pre obuke. Ukoliko je definisan rečnik, sve reči na koje se naiđe u korpusu za obuku, a koje se ne nalaze u rečniku, mogu biti ili ignorisane, ili modelovane kao jedna dodatna „reč“, koja se najčešće označava sa <*unk*>. Ukoliko se rečniku pridruži <*unk*>, odgovarajuće statistike se koriste pri korišćenju modela, za svaku nepoznatu reč na koju se naiđe.

ngram-merge

Ukoliko postoji potreba da se koristi više modela, koji su obučavani na korpusima različitih veličina i tipova sadržaja, pri čemu je pri interpolaciji dobijenih modela neophodno podesiti pojedinačne težine, ovaj izvršni program omogućava efikasnu realizaciju obuke krajnjeg modela.

ngram

Primena modela na nekom tekstu, ili prosto evaluacija modela na nekom skupu podataka za testiranje, izvodi se pomoću ovog izvršnog programa. Pored evaluacije standardnog modela računanjem perpleksnosti na skupu za testiranje

(pri čemu je moguće pristupiti detaljima proračuna perpleksnosti na nivou rečenice ili čak na nivou samih *N*-grama), implementirani su i različiti načini interpolacije modela. Osim standardnog modela, podržani su i klasni, faktorisani, keš (eng. *cache*) i mnogi drugi tipovi modela. O nekim od ovih specifičnih vrsta modela biće više reči u odeljku 2.1.4.

Na osnovu obučenih modela, ovim izvršnim programom moguće je generisati željeni broj rečenica (na slučajan način određenih dužina), što je veoma korisno pri analizi kvaliteta modela.

ngram-class

Ovaj izvršni program koristi se za automatsko izvođenje klasa reči na osnovu bigram statistika određenog korpusa. Detalji postupka automatske klasterizacije reči, koji je implementiran u okviru *SRILM toolkit*-a (automatska klasterizacija se, kao i grupisanje na osnovu manuelno implementiranih pravila, inače može vršiti na mnogo različitih načina), biće izloženi u prvom odeljku poglavlja 5, gde će biti prikazani rezultati poređenja automatske klasterizacije i grupisanja korišćenjem morfoloških informacija.

nbest-lattice

Procenom verovatnoće hipoteza i izborom najverovatnije, pri čemu se hipoteze mogu definisati manuelno, moguće je ovim izvršnim programom simulirati evaluaciju jezičkog modela u okviru sistema za automatsko prepoznavanje govora. Dakle, ukoliko je potrebno da se kvalitet jezičkog modela izrazi preko *WER*, to se može izvesti pomoću ovog programa.

segment

Pomoću ovog izvršnog programa moguće je iskoristiti jezički model radi segmentacije teksta na rečenice. Naravno, važno je voditi računa o tome da je za model koji će se koristiti u ovu svrhu potrebno koristiti korpus za obuku koji sadrži znake interpunkcije, koji se inače uklanjaju za potrebe modela jezika kao što su oni koji se koriste u okviru sistema za prepoznavanje govora. Istraživanje na ovu temu za srpski jezik biće izloženo u drugom odeljku poglavlja 5.

2.1.4 VRSTE *N*-GRAM MODELA I INTERPOLACIJA

U odeljku 2.1.2 bilo je reči o različitim tehnikama ublažavanja raspodele verovatnoće i različitim načinima kombinovanja modela višeg i nižeg reda. U okviru istraživanja koje je predmet ove disertacije, korišćeni su *back-off* modeli,

najčešće uz primenu *Good-Turing* ublažavanja, odnosno *Katz back-off* modeli. Pored različitih tehnika ublažavanja i kombinovanja modela višeg i nižeg reda, jezički modeli se mogu klasifikovati i na osnovu drugih kriterijuma, koji su se uglavnom nametnuli tokom razvoja modela za različite savremene primene.

U nastavku će biti ukratko opisani koncepti koji se, pored standardnog *N*-gram modela, najčešće koriste.

Klasni modeli

Klasni modeli predstavljaju jedan od osnovnih koncepata za tretiranje problema nedovoljne količine podataka za obuku (Whittaker & Woodland, 2001). Združivanjem reči koje se pojavljuju retko, ali u sličnim kontekstima, postiže se određen nivo generalizacije. Klase se mogu izvoditi na osnovu statistika korpusa za obuku, što je računarski veoma zahtevan proces, ili manuelno, što zahteva ekspertsko znanje. I u jednom i u drugom slučaju, problem predstavlja određivanje optimalnog broja klasa. U praksi se klasni modeli često koriste kao pomoćni modeli, odnosno vrši se interpolacija osnovnog i klasnog modela. Ipak, u nekim situacijama, kada postoje ograničenja po pitanju memorijskih kapaciteta ili procesorske moći uređaja, ili je jednostavno korpus za obuku izuzetno mali, klasni modeli se mogu koristiti i kao samostalni.

Keš modeli

Keš modeli se zasnivaju na pretpostavci da pojava određene reči tokom razgovora (ili u tekstu) povećava verovatnoću njene ponovne pojave u bliskoj budućnosti (odnosno, ako je u pitanju tekst, na poziciji nedaleko od prvog pojavljivanja). Ovakvi modeli pogodni su za automatsko prepoznavanje govora, ali samo u primenama ove govorne tehnologije u kojima se očekuje dijalog u vidu kratkih rečenica. Kao i klasni modeli, i ovakvi modeli se uglavnom koriste kao pomoćni modeli (Kuhn et al, 1990).

Diskriminativni modeli

Diskriminativni modeli se, za razliku od prethodno navedenih tipova modela, uglavnom koriste kao osnovni modeli. Oni su specifični po tome što se verovatnoće *N*-grama (ili brojevi pojavljivanja *N*-grama) modifikuju tako da bolje odgovaraju određenom zadatku, iako na taj način lošije reprezentuju korpus za obuku (Broman & Kurrimo, 2005). Modifikacija verovatnoća može da se izvodi i empirijski, na samo određenom broju *N*-grama (uz posledičnu normalizaciju svih verovatnoća). Ponekad se koriste i dodatni korpusi (sadržaja specifičnog za određenu primenu, koji može biti i više puta ponovljen isti tekst) koji se dodaju osnovnom korpusu, kako bi se pojedini *N*-grami istakli.

Struktuirani modeli

Kod struktuiranih modela rečenica se posmatra kao struktura stabla, koja je generisana od strane gramatike nezavisne od konteksta (eng. *context-free grammar*) (Nivre, 2005). U ovakvoj strukturi listovi predstavljaju pojedine reči (terminalne simbole), dok čvorovi predstavljaju neterminalne simbole, odnosno, određene sintaksne kategorije. Za konstrukciju stabla na osnovu date rečenice primenjuje se statistički pristup. S druge strane, za svaku novu rečenicu može se odrediti verovatnoća da ju je generisala data gramatika. Teorijski, ovakvi modeli mogu da modeluju značajno širi kontekst nego što to mogu standardni *N*-gram modeli. Takođe, oni su iz očiglednih razloga od posebnog interesa za lingviste. Međutim, njihova upotreba nije pogodna u aplikacijama kod kojih se očekuje spontaniji govor.

2.2 Jezički modeli bazirani na neuronskim mrežama

Razvoj veštačkih neuronskih mreža započet je polovinom prošlog veka, a sa razvojem različitih struktura i rešavanjem ograničenja koja su postojala kod prvih verzija neuronskih mreža, one su postajale sve interesantnija tema za istraživanje u različitim oblastima. Tako se danas neuronske mreže koriste u sistemima za automatsko prepoznavanje rukom pisanih simbola (eng. *Optical Character Recognition – OCR*), štampanih tekstova, zatim u ASR i TTS sistemima, kompjuterskoj viziji, mašinskom prevođenju i u rešavanju mnogih drugih aktuelnih problema.

U nastavku ovog odeljka predstavljene su strukture neuronskih mreža koje su značajne u oblasti obrade prirodnog jezika, opisan je standardni proces obuke NN, predstavljen je najpoznatiji alat za obuku NN modela baziranih na rekurentnim neuronskim mrežama (eng. *Recurrent Neural Networks – RNN*) – *RNNLM toolkit*, a zatim je dat i pregled aktuelnih istraživanja, odnosno poboljšanja i adaptacije osnovnih struktura NN za potrebe govornih i jezičkih tehnologija.

2.2.1 Osnovne strukture neuronskih mreža koje se koriste u modelovanju jezika

Prvi zapažen pokušaj da se jezik modeluje korišćenjem neuronske mreže vezuje se za ime Džef Elman. Elman je koristio rekurentnu neuronsku mrežu za modelovanje na osnovu rečenica koje su prethodno bile generisane pomoću gramatičkih pravila (Elman, 1990). Rad Holgera Švenka rezultovao je dokazima da se NNLM ponaša bolje od *N*-gram modela u okviru savremenih ASR sistema (Schwenk & Gauvain, 2005). Pored toga, Švenk je dokazao da NNLM sadrži

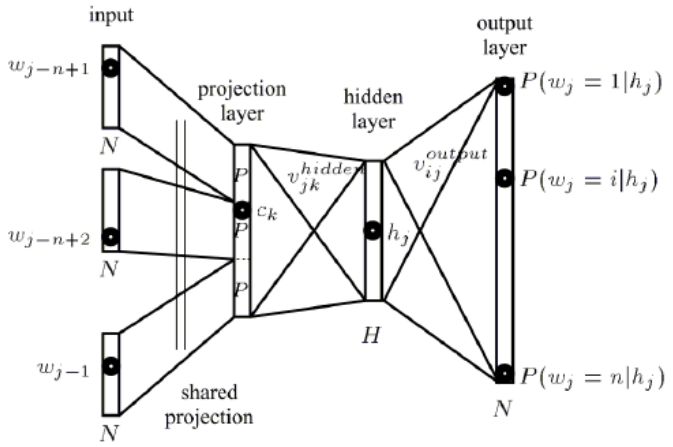
informacije koje su, u izvesnoj meri, komplementarne u odnosu na one koje pruža *N*-gram model te da ih u određenim primenama ima smisla kombinovati. Jošua Bendžio predložio je model neuronske mreže sa propagacijom unapred (eng. *Feed-Forward Neural Network – FFNN*) (Bengio et al, 2003), a neposredno nakon toga, Tomaš Mikolov je nastavio istraživanje u pravcu rekurentnih neuronskih mreža (RNN) (Mikolov et al, 2010). Najnovija istraživanja obuhvatila su i ispitivanja sa mrežama sa propagacijom uverenja (eng. *Deep Belief Network – DBN*) (Arisoy et al, 2012), kao i istraživanja mogućnosti uvođenja različitih dopunskih informacija pri obuci NNLM koje podrazumevaju morfološke, sintaksne, tematske (Mikolov & Zweig, 2012) i druge (Wu et al, 2012).

Neuronske mreže sa propagacijom unapred

Struktura neuronske mreže sa propagacijom unapred (Bengio et al, 2003) prikazana je na slici 1. Ovakav tip neuronske mreže koristi se ograničenim kontekstom, te su jezički modeli bazirani na ovakvim mrežama vrlo slični *N*-gram modelima. Na ulaz mreže dovodi se sekvenca koja čini istoriju dužine $N-1$. Svaka reč ove sekvence predstavljena je pomoću takozvanog „1-od-V" koda, gde je V veličina rečnika (na slici 1 veličina rečnika označena je sa N). Ovaj kôd podrazumeva da je svaka reč predstavljena vektorom dimenzije V, u kome svi elementi imaju vrednosti 0, osim elementa sa indeksom koji odgovara indeksu konkretne reči u rečniku, koji ima vrednost 1. Ova 1-od-V reprezentacija istorije (za neku reč w_j to je sekvenca reči $w_{j-N+1}, \dots, w_{j-1}$) zatim se linearnom projekcijom, korišćenjem zajedničke projekcione matrice $\boldsymbol{P}$, propagira ka projekcionom sloju, koji predstavlja prostor znatno manje dimenzije od dimenzije ulaza. S obzirom na to da je matrica $\boldsymbol{P}$ deljena među rečima koje čine sekvencu $w_{j-n+1}, \dots, w_{j-1}$, ona je identična u trenutku projekcije bilo koje od reči iz te sekvence.

Radi ilustracije dimenzionalnosti ulaznog i projekcionog sloja, uzmimo za primer modelovanje jezika pomoću konteksta dužine 4 (kvintagrami) uz veličinu rečnika N = 50.000. Na ulaznom sloju tada bi postojalo 200 hiljada binarnih promenljivih, od kojih bi samo 4 imale vrednost 1. Projekcija se često u praksi radi na prostor dimenzije 30 (ili nešto veće) po jednoj reči, što bi značilo da bi u datom primeru projekcioni sloj bio dimenzije 120. Nakon projekcionog sloja sledi tzv. skriveni sloj sa nelinearnom aktivacionom funkcijom (često se koristi tangens hiperbolična ili sigmoidalna funkcija). Za rečnik veličine 50 hiljada ovaj sloj bi, po nekim empirijskim informacijama, trebalo da bude dimenzije između 100 i 300. Poslednji sloj FFNN mreže je izlazni sloj, koji je dimenzije jednakoj veličini rečnika V (odnosno N na slici 1). Kada je NN obučena, ovaj sloj predstavlja raspodelu

verovatnoća $P(w_j|h_j)$, gde h_j predstavlja kontekst (istoriju) $w_{j-n+1}, \dots, w_{j-1}$. U opisanom primeru kvintagram NN modela, na izlazu bismo imali raspodelu verovatnoća $P(w_j|w_{j-1}, w_{j-2}, w_{j-3}, w_{j-4})$.



Slika 1. Struktura neuronske mreže sa propagacijom napred

U jednom od istraživanja Mikolov je predložio da se projekcija istorije u prostor niže dimenzionalnosti radi pomoću prethodno obučene NN za bigram model (Mikolov et al, 2011b). Obuka je u tom slučaju jednostavnija i lakša za razumevanje (a samim tim proces otklanjanja eventualne greške u kodu sistema oduzima mnogo manje vremena). Rezultati dobijeni na ovaj način gotovo su identični onima koji se dobijaju korišćenjem prvobitno predložene strukture.

Rekurentne neuronske mreže

Osnovna struktura RNN modela jezika opisana je u (Mikolov et al, 2010), a dorade, uglavnom vezane za efikasnost procesa obuke, opisane su u (Mikolov et al, 2011c). Osnovna razlika između FFNN i RNN strukture jeste u reprezentaciji istorije. Dok je kod FFNN strukture istorija i dalje ograničena na prethodnih nekoliko reči, kod RNN strukture do reprezentacije istorije se dolazi u procesu obuke i ona obuhvata teorijski sve prethodne reči. U skladu sa tim, modeli jezika bazirani na rekurentnim neuronskim mrežama mogu, bar u teoriji, uočiti i pravilnosti koje se manifestuju na veoma dugim sekvencama reči, čak i van granica pojedinih rečenica. Eksperimenti su potvrdili da RNN modeli jezika po uspešnosti modelovanja jasno prevazilaze mogućnosti FFNN modela (Mikolov, 2012).

Struktura modela jezika baziranog na RNN prikazana je na slici 2. Ovde se na ulazni sloj dovodi vektorska reprezentacija aktuelne reči, $\boldsymbol{w}(t)$ (koristi se, kao i kod FFNN, 1-od-*V* kodovanje). Pored ovog vektora, na ulazni sloj dovodi se i vektor $\boldsymbol{s}(t-1)$, koji predstavlja izlazne vrednosti skrivenog sloja NN iz prethodnog koraka. Nakon obuke ovakve mreže, na izlaznom sloju $\boldsymbol{y}(t)$ dobija se raspodela verovatnoća $P(w_{t+1}|w_t, \boldsymbol{s}(t-1))$. Ovde je važno napomenuti da rekurentne mreže predstavljaju posebnu vrstu rekurzivnih neuronskih mreža. Rekurzivne mreže, u opštem slučaju, operišu sa hijerarhijskim strukturama gde se reprezentacija potomka kombinuje sa reprezentacijom pretka, dok se kod rekurentnih mreža radi reprezentacije tekućeg trenutka u okviru linearne progresije vremena kombinuju prethodni trenutak i skrivena reprezentacija.

Obuka rekurentne neuronske mreže podrazumeva određivanje parametara matrica **U** i **W**, koje opisuju veze između ulaznog i skrivenog sloja, kao i matrice **V**, koja predstavlja vezu između skrivenog i izlaznog sloja. Obuka mreže može se izvesti pomoću algoritma propagacije unazad (eng. *backpropagation – BP*) (Boden, 2002) ili algoritma propagacije unazad kroz vreme (eng. *backpropagation through time – BPTT*) (Bengio et al, 1994). Izlazi skrivenog i izlaznog sloja mreže računaju se pomoću izraza (2.30) i (2.31), respektivno.

$$s_j(t) = f\left(\sum_i w_i(t)u_{ji} + \sum_l s_l(t-1)w_{jl}\right) \tag{2.30}$$

$$y_k(t) = g\left(\sum_j s_j(t)v_{kj}\right) \tag{2.31}$$

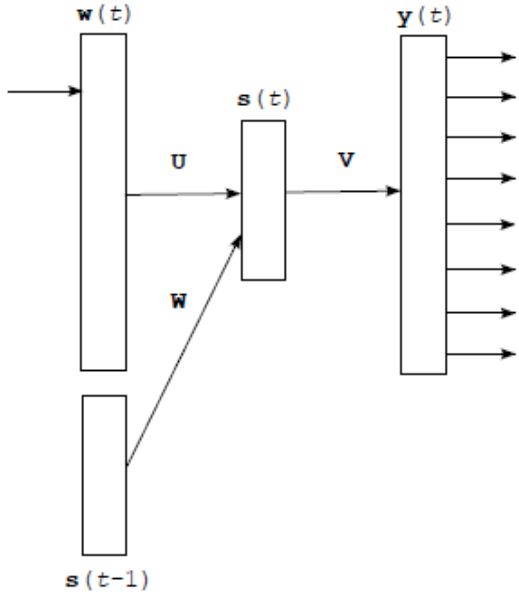
U navedenim izrazima *f* i *g* označavaju *sigmoid* (2.32) i *softmax* (2.33) funkcije.

$$f(z) = \frac{1}{1+e^{-z}} \tag{2.32}$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{2.33}$$

Funkcija data izrazom (2.33) koristi se na izlaznom sloju kako bi se osiguralo da izlazi formiraju validnu raspodelu verovatnoće, u smislu da su sve vrednosti veće od 0 i da je njihova suma jednaka 1.

Primer modela jezika baziranog na RNN dat je u prilogu 2. Model je obučen pomoću alata *RNNLM toolkit* (Mikolov et al, 2011d), o kome će biti više reči u odeljku 2.2.3.



Slika 2. Struktura rekurentne neuronske mreže

2.2.2 ALGORITMI ZA OBUKU NN MODELA JEZIKA

Neuronske mreže različitih struktura obučavaju se pomoću tzv. algoritma propagacije (gradijenta greške) unazad kroz vreme (eng. *backpropagation through time – BPTT*), koji se oslanja na tzv. metodu stohastičkog opadanja gradijenta (eng. *Stochastic Gradient Descent – SGD*). Ovaj algoritam podrazumeva propagaciju gradijenta greške u celokupnoj mreži unazad kroz vreme pomoću rekurentnih težina, što omogućava modelu da „uhvati" korisne informacije koje su sadržane u skrivenom sloju. U nastavku će biti izložen osnovni *BP*, a zatim i *BPTT* algoritam.

BP algoritam

Pomoću metode stohastičkog opadanja gradijenta, težinski koeficijenti matrica mreže preračunavaju se nakon svakog uzorka. Za dobijanje gradijenta vektora greške na izlaznom sloju koristi se tzv. kriterijum unakrsne entropije (eng. *cross entropy criterion*). Gradijent vektora greške se propagira sa izlaznog unazad ka skrivenom sloju, a zatim, putem rekurentnih težinskih koeficijenata, ka ulaznom sloju, odnosno „unazad u vremenu". Tokom obuke se koristi validacioni korpus

kako bi se obezbedila mogućnost ranog zaustavljanja algoritma (da ne bi došlo do preteranog prilagođavanja podacima za obuku), kao i za kontrolisanje parametra koji opisuje brzinu učenja. Tokom obuke, iterira se kroz kompletan korpus za obuku u nekoliko tzv. epoha (eng. *epochs*). Do konvergencije algoritma uglavnom dolazi nakon 8-20 iteracija. Što se tiče brzine učenja, ona se u početku postavlja na vrednost $\alpha = 0{,}1$. Ova vrednost se koristi dokle god postoji značajno poboljšanje na validacionim podacima. U eksperimentima opisanim u (Mikolov, 2012), značajnim poboljšanjem smatra se opadanje entropije za makar 0,3%. Kada poboljšanje padne ispod definisanog praga, u svakoj sledećoj epohi brzina učenja se prepolovi i obuka se nastavlja dok poboljšanje ponovo ne padne ispod definisanog praga i tada je obuka završena. Broj epoha sa početnom i umanjenim vrednostima brzine učenja može se i unapred definisati, te se obuka može izvesti i bez validacionog skupa podataka.

Matrice **U**, **V** i **W** inicijalizuju se slučajnim malim vrednostima. U eksperimentima Mikolova ovi brojevi se dobijaju na osnovu Gausove raspodele sa srednjom vrednošću 0 i varijansom 0,1. Jedna epoha obuke RNN izvodi se u sledećih 8 koraka:

1) Brojač „vremena“ se postavlja na $t = 0$, a stanja neurona skrivenog sloja $s(t)$ na 1.

2) Brojač se inkrementuje za 1.

3) Na ulazni sloj dovodi se vektorska reprezentacija $\boldsymbol{w}(t)$ reči w_t.

4) Stanje skrivenog sloja $\boldsymbol{s}(t-1)$ se kopira na ulazni sloj.

5) Vrši se propagacija unapred kako bi se dobili vektori $\boldsymbol{s}(t)$ i $\boldsymbol{y}(t)$.

6) Računa se gradijent greške $\boldsymbol{e}(t)$ na izlaznom sloju.

7) Greška se propagira unazad kroz mrežu, a težinski koeficijenti matrica se pritom menjaju po potrebi.

8) Ako nisu obrađeni svi podaci, prelazi se na korak 2.

Cilj obuke NN zapravo je maksimizacija verodostojnosti podataka za obuku λ, što je opisano sledećim izrazom:

$$f(\lambda) = \sum_{t=1}^{T} \log y_{l_t}(t). \tag{2.34}$$

U ovom izrazu, podaci za obuku (reči) označeni su sa t, a l_t predstavlja indeks ispravno „predviđene“ reči za t-ti uzorak. Gradijent vektora greške na izlaznom sloju $\boldsymbol{e}_o(t)$ računa se pomoću kriterijuma unakrsne entropije, čiji je cilj da se

maksimizuje izglednost korektne klase, a računa se kao što je prikazano sledećim izrazom:

$$e_o(t) = d(t) - y(t). \tag{2.35}$$

U izrazu (2.35) vektor $\boldsymbol{d}(t)$ predstavlja ciljni vektor koji predstavlja reč w_{t+1} (sa reprezentacijom $\boldsymbol{w}(t+1)$ na ulaznom sloju) koja je trebala biti predviđena ($\boldsymbol{d}(t)$ je, dakle, 1-od-V kôd reči w_{t+1}). Ovde je važno napomenuti da je potrebno koristiti upravo kriterijum unakrsne entropije, a ne kriterijum srednje kvadratne greške (eng. *mean square error – MSE*). Iako bi mreža funkcionisala i korišćenjem MSE kriterijuma, rezultat bi bio suboptimalan. Težine matrice **V** između skrivenog $\boldsymbol{s}(t)$ i izlaznog $\boldsymbol{y}(t)$ sloja koriguju se na način prikazan izrazom (2.36):

$$v_{jk}(t+1) = v_{jk}(t) + s_j(t)e_{ok}(t)\alpha. \tag{2.36}$$

U datom izrazu α predstavlja brzinu učenja, j predstavlja iterator kroz skriveni, a k kroz izlazni sloj mreže, $s_j(t)$ je izlaz j-tog neurona skrivenog sloja, a $e_{ok}(t)$ je gradijent greške k-tog neurona izlaznog sloja. Ako se koristi L2 regularizacija, težine se koriguju pomoću izraza (2.37), u kome je β regularizacioni parametar:

$$v_{jk}(t+1) = v_{jk}(t) + s_j(t)e_{ok}(t)\alpha - v_{jk}(t)\beta. \tag{2.37}$$

Regularizacija se vrši da bi se vrednosti težina održale malima (radi kompaktne reprezentacije). Izraz (2.37) može se predstaviti i u matričnom obliku, kao u sledećem izrazu:

$$V(t+1) = V(t) + s(t)e_o(t)^T\alpha - V(t)\beta. \tag{2.38}$$

Nakon korekcije matrice **V**, gradijent greške se propagira na skriveni sloj (2.39).

$$e_h(t) = d_h(e_o(t)^T V, t) \tag{2.39}$$

U izrazu (2.39), funkcija $d_h()$ primenjuje se na svaki pojedinačni element i predstavljena je izrazom (2.40):

$$d_{hj}(x,t) = xs_j(t)\Big(1 - s_j(t)\Big). \tag{2.40}$$

Težinski koeficijenti matrice **U** (između ulaznog $\boldsymbol{w}(t)$ i skrivenog $\boldsymbol{s}(t)$ sloja) u nastavku se koriguju na način prikazan izrazom (2.41):

$$u_{ij}(t+1) = u_{ij}(t) + w_i(t)e_{hj}(t)\alpha - u_{ij}(t)\beta. \tag{2.41}$$

Ovde je važno napomenuti da je na ulaznom sloju u svakom trenutku aktivan tačno 1 neuron. Kao što se vidi iz izraza (2.41), težinski koeficijenti vezani za ostale

neurone ne menjaju se, tako da se postupak korekcije težina može ubrzati preračunavanjem samo onih težinskih koeficijenata koji odgovaraju aktivnom neuronu.

Rekurentni težinski koeficijenti matrice **W** koriguju se pomoću sledećeg izraza:

$$w_{lj}(t+1) = w_{lj}(t) + s_l(t-1)e_{hj}(t)\alpha - w_{lj}(t)\beta. \quad (2.42)$$

BPTT algoritam

Iako *BP* algoritam podrazumeva optimizaciju predikcije reči pomoću prethodne reči i stanja skrivenog sloja korekcijom težinskih matrica, nikakve korisne informacije koje se pri tome dobijaju ne smeštaju se u sam skriveni sloj.

Algoritam *BPTT* predstavlja nadogradnju osnovnog *BP* algoritma. Zasniva se na ideji da RNN posmatrana u *N* vremenskih koraka može da se posmatra i kao FFNN sa *N* skrivenih slojeva (tzv. duboka ili dubinska struktura), što je prikazano na slici 3. Takva struktura može biti obučavana pomoću standardne metode opadanja gradijenta (eng. *gradient descent*). Pri tome se greške propagiraju sa skrivenog sloja $\boldsymbol{s}(t)$ ka skrivenom sloju koji odgovara prethodnom trenutku $\boldsymbol{s}(t-1)$, a matrica **W** biva korigovana. Ovaj algoritam zahteva da se čuvaju stanja skrivenog sloja iz svih prethodnih vremenskih koraka.

Propagacija greške vrši se rekurzivno, kao što je prikazano sledećim izrazom:

$$e_h(t-\tau-1) = d_h(e_h(t-\tau)^T W, t-\tau-1). \quad (2.43)$$

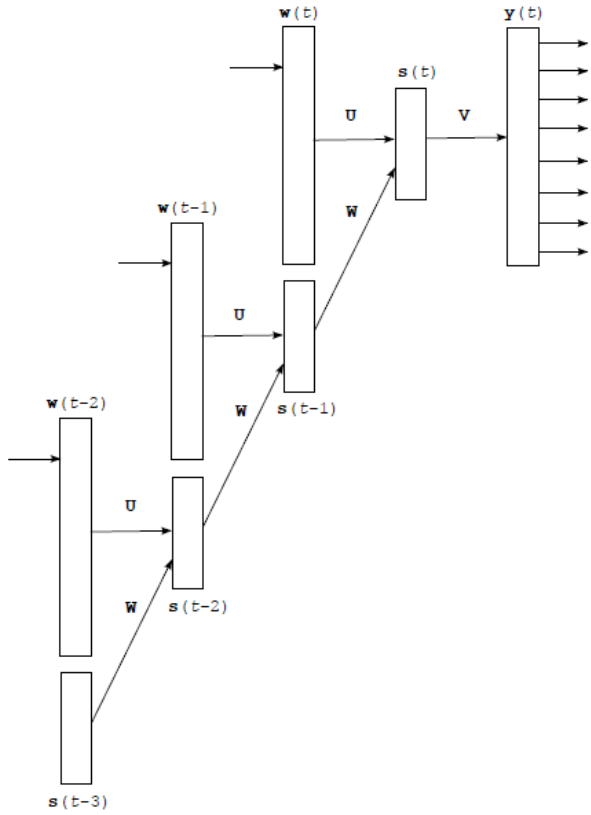
Opisani proces „odmotavanja“ RNN u vremenu može se, teorijski, izvoditi u onoliko koraka koliko je podataka (reči) do datog trenutka obrađeno. U praksi, gradijenti brzo nestaju (a može doći i do „eksplozije“ tih vrednosti, što je veoma retko), zbog čega se odmotavanje uglavnom vrši u svega nekoliko koraka (oko 5). Empirijski je utvrđeno da *BP* algoritmom RNN može da se obuči tako da dobro modeluje jezička pravila koja se protežu na kontekste do 4-5 reči, dok *BPTT* ovaj kontekst proširuje na 8-9 (Mikolov & Zweig, 2012).

Pri *BPTT* obuci, težinski koeficijenti matrica **U** i **W** koriguju se kao što je dato izrazima (2.44) i (2.45), respektivno. U ovim izrazima, *T* označava broj koraka odmotavanja u vremenu.

$$u_{ij}(t+1) = u_{ij}(t) + \sum_{z=0}^{T} w_i(t-z)e_{hj}(t-z)\,\alpha - u_{ij}(t)\beta \quad (2.44)$$

$$w_{lj}(t+1) = w_{lj}(t) + \sum_{z=0}^{T} s_l(t-z-1)e_{hj}(t-z)\alpha - w_{lj}(t)\beta \qquad (2.45)$$

Važno je napomenuti da se korekcija matrice **U** mora vršiti odjednom, a ne inkrementalno nakon svakog koraka odmotavanja, jer to dovodi do nestabilnosti pri obuci.



Slika 3. RNN posmatrana u vremenu kao FFNN mreža, u konkretnom primeru 3 koraka unazad u vremenu

2.2.3 *RNNLM TOOLKIT* – ALAT ZA OBUKU RNN MODELA JEZIKA

Za potrebe kreiranja jezičkih modela baziranih na neuronskim mrežama najpoznatiji je alat *RNNLM toolkit* (Mikolov et al, 2011d), koji je javno dostupan i koji služi kao osnova svim aktuelnim istraživanjima vezanim za NNLM. U ovom odeljku će biti samo ukratko opisane najvažnije mogućnosti koje ovaj alat nudi.

Aktuelna verzija *RNNLM toolkit*-a pruža mogućnost obučavanja modela na osnovu rekurentne neuronske mreže i uključuje sva ubrzanja obuke koja su razvijena u proteklih nekoliko godina. Osnovno ubrzanje obuke postignuto je tzv. faktorisanjem izlaznog sloja, o čemu će biti više reči u odeljku 2.2.4. Osim obučavanja RNN modela jezika, moguće je generisati rečenice pomoću gotovog modela. Na ovaj način pruža se mogućnost kreiranja *N*-gram modela u ARPA formatu pomoću alata *SRILM*, koji aproksimira dati RNN model. Naime, potrebno je obučiti RNN model, a zatim generisati proizvoljno velik broj rečenica pomoću tog modela. Te rečenice čine korpus za obuku *N*-gram modela pomoću *SRILM*-a. Ova mogućnost je od velike koristi za ASR sisteme kod kojih nije jednostavno adaptirati kôd na nov format modela jezika. Ono što je mnogo važnije za istraživanja, *RNNLM toolkit* pruža mogućnost interpolacije *N*-gram i NN modela jezika. Pokazano je da se kombinacijom ova dva tipa modela dobijaju uočljiva poboljšanja po pitanju *ppl*, kao i *WER*.

Radi ilustracije procesa obuke i evaluacije RNN modela, u nastavku je dat primer poziva opisanog alata iz komandne linije za obuku i test, respektivno.

```
RNNLM.exe -train train -valid valid -rnnlm model -hidden
     70 -rand-seed 1 -debug 2 -bptt 3 -class 200
RNNLM.exe -rnnlm model -test test
```

Osim činjenice da izvorni kôd nije zavisan od *third-party* biblioteka, što ga čini portabilnim i korisnim u smislu integracije u različite sisteme, svi parametri vezani za obuku modela ostavljeni su „vidljivim“ i na raspolaganju su za eksperimentisanje korisnicima izvršne verzije programa. Dakle, eksperimenti na novim jezicima (sa postojećim strukturama NN) mogu se raditi i bez poznavanja koda. U datom primeru poziva programa za obuku vidi se da je moguće menjati veličinu, odnosno broj neurona skrivenog sloja (*-hidden*), broj klasa kod faktorisanog izlaza (*-class*), broj koraka propagacije greške unazad u vremenu (*-bptt*). Pri tome, problem „nestajućeg gradijenta“ pri *BPTT* obuci je donekle rešen reskaliranjem vrednosti tokom obuke. Međutim, ovo su samo neke od mogućnosti koje ovaj alat pruža. Primera radi, kao ulaz modelu se može zadati *N*-best lista dobijena pomoću ASR sistema. Model se može iskoristiti za ocenjivanje svih hipoteza i izbor najverovatnije. Na ovaj način se, na primer, može simulirati procena doprinosa modela redukciji *WER* kod sistema za automatsko prepoznavanje govora. Primer sadržaja modela jezika koji se može dobiti pomoću *RNNLM toolkit*-a prikazan je u prilogu 2.

2.2.4 AKTUELNA ISTRAŽIVANJA U OBLASTI MODELOVANJA JEZIKA POMOĆU NEURONSKIH MREŽA

Od početka korišćenja neuronskih mreža u modelovanju jezika razvoj je uglavnom bio orijentisan ka optimizaciji brzine obuke modela, s obzirom na to da je obuka neuronskih mreža računarski zahtevan postupak i da, kada se radi sa ogromnim korpusima i na velikim rečnicima, može da traje i po nekoliko dana. Pored toga, nekoliko istraživanja bilo je usmereno na uvođenje sintaksnih i semantičkih informacija u vidu dodatnih obeležja reči na ulaznom sloju neuronske mreže. Najvažnija od ovih istraživanja biće izložena u nastavku ovog odeljka.

Faktorisan RNN model jezika

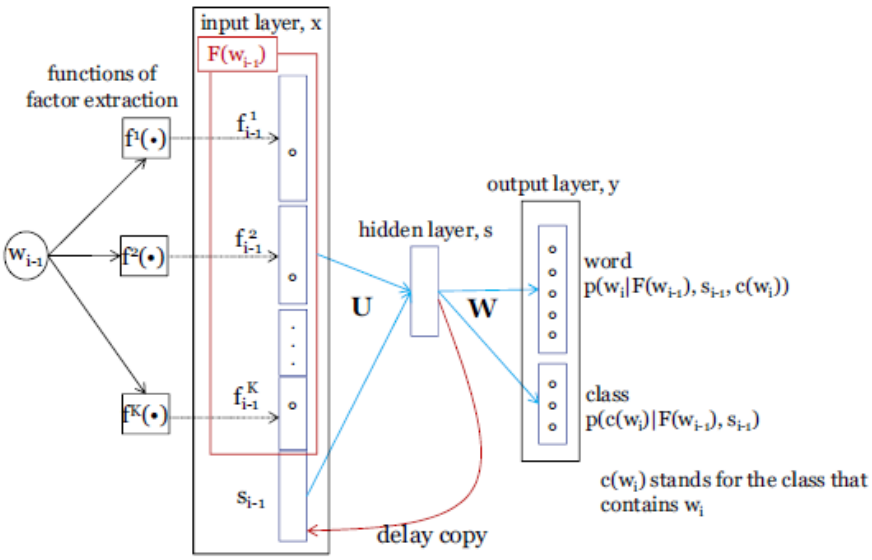
Faktorisan model jezika (fRNNLM) predstavlja nadogradnju RNN modela uvođenjem dodatnih morfoloških faktora (obeležja) u proces obuke mreže putem eksplicitne integracije (Wu et al, 2012).

U prethodnom odeljku pomenuto je faktorisanje izlaznog sloja mreže radi ubrzanja procesa obuke. Iako se koristi isti termin, faktorisan model jezika i faktorisan izlazni sloj ne odnose se na isti pojam. Ovo se može razjasniti analizom strukture faktorisanog modela jezika koja je prikazana na slici 4 (Wu et al, 2012).

Pojam faktorisanog modela odnosi se na izdvajanje (ekstrakciju) morfoloških (ili nekih drugih) obeležja vezanih za pojedine reči koje se dovode na ulaz mreže. Ovakav model, umesto predikcije verovatnoća $P(w_i|w_{i-1}, s_{i-1})$, vrši predikciju verovatnoća $P(w_i|F(w_{i-1}), s_{i-1})$. Ove verovatnoće eksplicitno zavise od skupa koji sadrži K faktora vezanih za prethodnu reč, a implicitno zavise i od faktora vezanih za sve ostale prethodne reči zbog rekurzivnog elementa (na slici je rekurzivna veza označena sa *delayed copy*). Vektor $F(w_{i-1})$ se dobija konkatenacijom K vektora faktora $f_{i-1}^k (k = 1, \dots, K)$, pri čemu je f_{i-1}^k oznaka za k-ti vektor faktora prethodne reči w_{i-1}. Funkcije koje se koriste za ekstrakciju faktora označene su na slici 4 sa $f^k(\cdot)$. Faktori (obeležja) reči mogu biti, na primer, morfološke kategorije ili koreni reči, kao i neke druge lingvističke informacije. Kodovanje obeležja identično je kodovanju reči pri vektorskoj reprezentaciji u standardnom RNN modelu (dakle, 1-od-V, pri čemu je V različito za svako obeležje). Ulazni sloj formira se na osnovu vektora $F(w_{i-1})$ i kopije skrivenog sloja iz prethodnog vremenskog koraka, kao što je prikazano izrazom (2.46).

$$x_i = [F(w_{i-1}), s_{i-1}] \tag{2.46}$$

Matrica **U** u ovoj strukturi sadrži težinske koeficijente pojedinih faktora. Ostatak strukture odgovara standardnom RNN modelu.



Slika 4. Struktura faktorisanog RNN modela jezika

Pod faktorisanjem izlaznog sloja modela podrazumeva se svrstavanje reči u klase, pri čemu se polazi od pretpostavke da svaka reč pripada tačno jednoj klasi (Mikolov et al, 2011c). Izlazni sloj se pritom deli na dve celine. Prva celina predstavlja estimaciju raspodele *a posteriori* verovatnoće po svim klasama reči. Druga celina predstavlja estimaciju raspodele *a posteriori* verovatnoće svih reči koje pripadaju određenoj klasi. Konačno, tražena verovatnoća $P(w_i|F(w_{i-1}), s_{i-1})$ dobija se kao proizvod pomenute dve verovatnoće, kao što je prikazano izrazom (2.47).

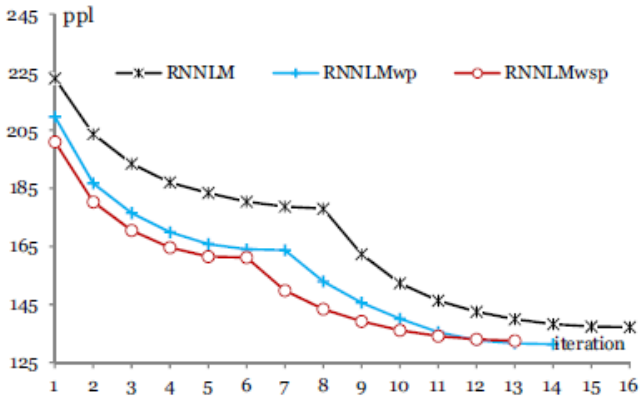
$$P(w_i|F(w_{i-1}), s_{i-1}) = P(c(w_i)|F(w_{i-1}), s_{i-1}) \times P\big(w_i|F(w_{i-1}), s_{i-1}, c(w_i)\big) \quad (2.47)$$

Ova arhitektura, koja podrazumeva razdvajanje izlaznog sloja na dva dela može u velikoj meri da ubrza proces obuke uz praktično neznatan gubitak na kvalitetu modelovanja jezika.

Faktorisan model jezika u istraživanju publikovanom u (Wu et al, 2012) testiran je na korpusu za engleski jezik. Međutim, za visoko inflektivne jezike poput češkog, arapskog, ruskog ili srpskog, uvođenje morfoloških informacija trebalo bi da dovede do još uočljivijih poboljšanja.

U konkretnom eksperimentu, za engleski jezik korišćen je korpus koji sadrži preko 30 miliona reči, pri čemu je veličina rečnika preko 150 hiljada. Veličina skrivenog sloja postavljena je na 480, a broj klasa na izlaznom sloju na 300. Istraživanje je pokazalo da fRNNLM uvodi poboljšanje u odnosu na RNNLM po pitanju redukcije *WER* za 0,4-0,8%, što je značajno, s obzirom na veoma dobre performanse osnovnog RNN modela. Naravno, oba poređena modela su interpolirana sa *Knesser-Ney N*-gram modelima kako bi se postigli što bolji rezultati i kako bi se videlo da fRNNLM zaista doprinosi tačnosti prepoznavanja govora novim informacijama koje uvodi u obuku.

Što se tiče (prostorne) kompleksnosti faktorisanog modela u odnosu na osnovni RNNLM, fRNNLM ima tačno $(|f^1| + \cdots + |f^K| - |V|) \times \mathrm{H}$ više slobodnih parametara nego RNNLM, koji ih ima $(|V| + H) \times \mathrm{H} + \mathrm{H} \times (C + |V|)$. Vremenska (računska) kompleksnost RNNLM modela je $(1 + H) \times \mathrm{H} \times \tau + \mathrm{H} \times |V|$, dok fRNNLM ima dodatnu kompleksnost $(K - 1) \times \mathrm{H} \times \tau$. U navedenim izrazima τ predstavlja broj koraka propagacije gradijenata grešaka u vremenu. Ovaj parametar se obično postavlja na vrednost 4 ili 5 te je $H \times |V|$ dominantan element u odnosu na $(K - 1) \times \mathrm{H} \times \tau$ pa se razlika u vremenskoj kompleksnosti RNNLM i fRNNLM može zanemariti. Štaviše, u paralelnom testiranju pokazalo se da fRNNLM brže konvergira od osnovnog modela (Wu et al, 2012), što je prikazano na slici 5.



Slika 5. Krive konvergencije RNNLM, faktorisanog modela koji se koristi oznakama vrste reči (RNNLMwp) i faktorisanog modela koji se koristi oznakama vrste reči i korenima reči (RNNLMwsp)

Kontekstno-zavisan RNNLM

Slično konceptu faktorisanog modela, kontekstno-zavisni model predstavlja uvođenje dodatnih informacija na ulazni sloj neuronske mreže, kako bi se postigli bolji rezultati obuke (Mikolov & Zweig, 2012). U ovom slučaju, uz svaku reč koja se dovodi na ulaz neuronske mreže dovodi se i informacija o tome sa kojom verovatnoćom data reč pripada kojoj „temi" (eng. *topic*). Naime, svaki dokument se posmatra kao mešavina nekoliko tema. Svaka reč iz određenog dokumenta pripada pojedinim temama sa određenom verovatnoćom. Na osnovu konteksta, može se odrediti raspodela verovatnoća po temama za svaku reč i ta raspodela se pridružuje datoj reči kao ulazna informacija za NN.

Struktura kontekstno zavisnog modela, koji se od osnovnog RNNLM razlikuje po postojanju dodatnog sloja obeležja za teme (sa pratećim matricama težina **F** i **G**), prikazana je na slici 6. Obuka ovakve strukture analogna je osnovnoj *BPTT* obuci.

Za modelovanje tematskog konteksta koristi se postupak pod nazivom latentna Dirihleova alokacija (eng. *Latent Dirichlet Allocation – LDA*) (Blei et al, 2003). Ova procedura obuhvata posmatranje teksta na vrlo jednostavan način, koji podrazumeva zanemarivanje redosleda reči i vođenje računa samo o tome koje se reči pojavljuju u tekstu (eng. *bag-of-words*), i preslikavanje takve reprezentacije dokumenta u vektor male dimenzije, koji se tumači kao tematska reprezentacija. Primarna pretpostavka je da su teme po dokumentu zastupljene sa verovatnoćama koje čine Dirihleovu raspodelu. Svakoj temi koja je prisutna u dokumentu odgovara posebna raspodela verovatnoća po unigramima (rečima). Ključni parametar LDA je α, kojim se određuje oblik početne raspodele verovatnoća tema po dokumentu. Ako je $\alpha < 1$, raspodela ima izražen „vrh", što znači da će jednoj temi biti pridružena velika verovatnoća. Za $\alpha = 1$ raspodela je, zapravo, uniformna. Ukoliko je $\alpha > 1$, penalizuju se raspodele koje favorizuju bilo koju konkretnu temu. Rezultat LDA obuke je utvrđena vrednost α, kao i skup raspodela verovatnoća reči po temama β.

U istraživanju opisanom u (Mikolov & Zweig, 2012) korišćen je prozor prethodnih reči fiksne dužine za određivanje raspodele verovatnoće po temama. Zbog ovoga, potrebno je za svaku reč korigovati kontekstni vektor, što je u opštem slučaju složen i vremenski veoma zahtevan proces. U okviru opisanog istraživanja razvijen je i efikasniji način izračunavanja raspodele verovatnoća po temama za blok reči na osnovu skupa raspodela verovatnoća tema za pojedine reči. Naime, zapaženo je da je moguće doći do dobre aproksimacije raspodele verovatnoća po temama za blok reči prostim množenjem raspodela verovatnoća po temama za pojedinačne reči iz datog bloka, uz neophodnu normalizaciju rezultujuće raspodele, kao što je prikazano izrazom:

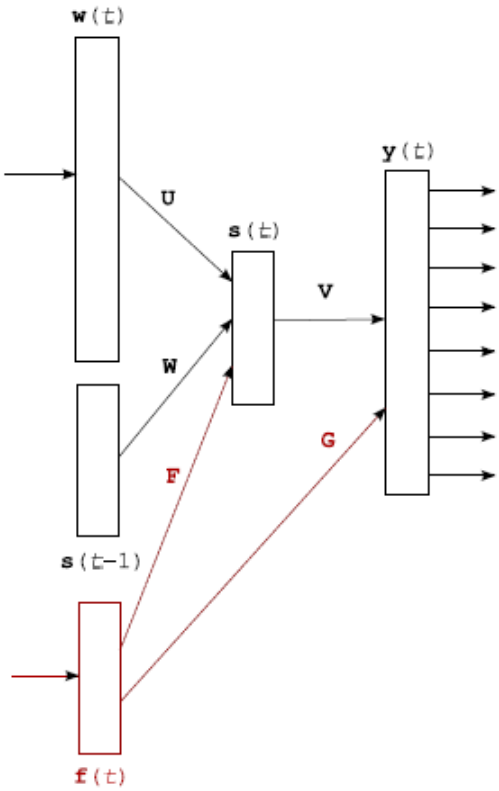
$$f(t) = \frac{1}{Z}\prod_{i=0}^{K} t_{w(t-i)}, \tag{2.48}$$

gde je $\boldsymbol{t}_{w(t)}$ je vektor koji predstavlja LDA raspodelu verovatnoća po temama za reč $w(t)$, a Z je normalizacioni koeficijent. Da bi ova aproksimacija bila uspešna, potrebno je izvršiti ublažavanje β matrice dodavanjem malih konstanti, kako bi se izbegle ekstremno male vrednosti verovatnoće. Poboljšanje načina izračunavanja $f(t)$ opisanog izrazom (2.48) može se postići davanjem različitih težina pojedinim rečima i to manjih težina onima koje su se pojavile u daljoj prošlosti. Jedna od mogućnosti opisana je izrazom:

$$f(t) = \frac{1}{Z} f(t-1)^{\gamma} t_{w(t)}^{(1-\gamma)}, \tag{2.49}$$

pri čemu parametar γ kontroliše brzinu adaptacije na nove teme, odnosno brzinu promene vektora tematskih obeležja. Vrednosti blizu 1 dopuštaju sporu adaptaciju na nove teme, dok male vrednosti omogućavaju brze promene vektora tematskih obeležja.

Pomenuto istraživanje dalo je rezultate u vidu redukcije *WER* za nekoliko ispitanih konfiguracija, međutim, u poređenju sa *state-of-the-art* modelom poboljšanje je skromno, uglavnom oko 0,1-0,2%. Rađene su optimizacije parametara modela vezanih za sloj obeležja i došlo se do zaključka da se najbolji rezultati postižu za veličinu tematskog sloja između 10 i 40, kada se pri tom koristi kontekst od prethodnih 50 reči, a parametar γ postavi na vrednost 0,95. Međutim, ovi parametri, iako mogu poslužiti kao orijentir, nemaju velikog značaja za istraživanja koja treba sprovesti na drugim jezicima. Štaviše, kontekst od 50 reči ima smisla koristiti na *Penn Treebank* korpusu (koji je korišćen za potrebe opisanog istraživanja), jer su susedne rečenice često u semantičkoj vezi, ali na nekom korpusu koji sadrži spisak potpuno nepovezanih rečenica, taj kontekst očigledno mora biti manji (Marcus et al, 1993).



Slika 6. Struktura kontekstno-zavisnog RNN modela jezika

NN model jezika sa struktuiranim izlaznim slojem

Izlazni sloj neuronske mreže predstavlja „usko grlo“ po pitanju brzine kod obuke modela jezika. Naime, korišćenje *softmax* funkcije pri računanju verovatnoća zahteva operaciju sumiranja po celom izlaznom sloju, koji je dimenzije jednake veličini rečnika, i to za svaku reč iz skupa za obuku mora da se ponovi. Ovo čini proces obuke NN veoma neefikasnim i često se rešava redukcijom skupa izlaznih verovatnoća koje se estimiraju. Naime, definiše se skup najčešćih reči u korpusu za obuku, i za njih se računaju izlazne verovatnoće. Za ostale reči koristi se pozadinski *N*-gram model. Iako ovo funkcioniše bolje nego da se koristi samo *N*-gram model, eliminacijom manje čestih reči iz izlaznog sloja NN predstavlja ograničenje kojim se gubi veliki deo informacija u kojima zapravo i leži ključna prednost NN u odnosu na *N*-gram modele. Pokazano je da se NNLM mnogo bolje ponašaju od *N*-gram

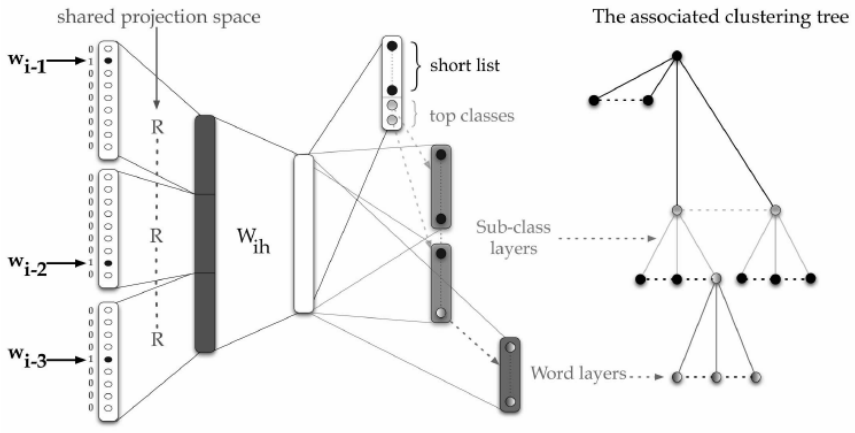
modela u slučajevima reči koje se retko pojavljuju u korpusu za obuku, dok za najčešće reči ta prednost nije toliko izražena (Oparin et al, 2012). Ovo je bila motivacija za definisanje strukture izlaznog sloja koja će biti prihvatljiva po pitanju efikasnosti obuke, a koja će ipak sadržati procene *a posteriori* verovatnoća za sve reči ulaznog rečnika. Prvi rezultati ovog istraživanja objavljeni su u (Le et al, 2011a) i (Le et al, 2011b), a noviji rezultati nakon sitnijih izmena i primene modela u okviru ASR sistema dati su u (Le et al, 2013).

Ovaj koncept klasnog NNLM nazvan je modelom jezika zasnovanim na neuronskoj mreži sa struktuiranim izlaznim slojem (eng. *Structured Output Layer – SOUL*). U SOUL NNLM izlazni rečnik je struktuiran pomoću klasifikacionog (klasterizacionog) stabla, pri čemu svaka reč pripada samo jednoj klasi i samim tim se svrstava u jedan od listova stabla. Ako je w_i i-ta reč u rečenici, sa $c_{1:D}(w_i) = c_1, \dots, c_D$ označena je putanja u klasifikacionom stablu koja vodi do date reči, pri čemu D predstavlja dubinu stabla. Sa $c_d(w_i)$ označena je klasa (potklasa) kojoj je dodeljena reč w_i, a $c_D(w_i)$ označava list stabla koji sadrži tu reč. Imajući u vidu SOUL strukturu izlaznog sloja, N-gram verovatnoća reči w_i u kontekstu h može se estimirati koristeći izraz sledeći izraz:

$$P(w_i|h) = P(c_1(w_i)|h) \prod_{d=2}^{D} P(c_d(w_i)|h, c_{1:d-1}). \tag{2.50}$$

Dakle, verovatnoća reči koja je estimirana na izlaznom sloju NN nije uslovljena samo nizom prethodnih reči h, već i nizom nebinarnih vrednosti kojima se koduje putanja do date reči kroz klasifikaciono stablo. Kao što je ranije rečeno, svaka reč pripada tačno jednoj krajnjoj klasi (listu stabla) te je moguće primeniti *softmax* funkciju na svakom nivou ove hijerarhijske reprezentacije. Struktura SOUL NNLM prikazana je na slici 7.

Vidi se da je struktura SOUL modela do skrivenog sloja identična strukturi standardnog NNLM. Izlazni sloj se, međutim, sastoji iz tri nivoa izlaza. Na prvom nivou direktno se estimiraju verovatnoće za reči iz liste najčešćih u korpusu za obuku (eng. *shortlist*), kao i verovatnoće određenog broja osnovnih klasa *OOS* (*out-of-shortlist*) reči, odnosno reči koje se ne nalaze na listi najčešćih. Osnovne klase definisane na prvom nivou izlaza služe kao koreni stabala pri daljoj klasifikaciji. Međuslojevi koji pri tome nastaju spadaju u drugi nivo izlaza. Ovaj nivo ne produkuje izlazne verovatnoće, već samo predstavlja međukorak pri klasifikaciji. Treći i poslednji nivo izlaza predstavlja nivo reči. Izlazne verovatnoće ovog nivoa predstavljaju estimacije uslovnih verovatnoća reči koje se ne nalaze na listi najčešćih. Uzevši u obzir izraz (2.50), svrha prvog, drugog i trećeg nivoa izlaza jeste određivanje verovatnoća $P(c_1(w_i)|h)$, $P(c_d(w_i)|h, c_{1:d-1})$ za $1 < d < D$ i $P(c_D(w_i)|h, c_{1:D-1})$, respektivno.



Slika 7. Struktura SOUL NN modela jezika

Pri obuci ovakvog sistema pored određivanja parametara NN potrebno je odrediti i strukturu klasifikacionog stabla, odnosno izlaznog rečnika. Postoji više načina da se izvrši klasifikacija reči. U okviru istraživanja opisanog u (Le et al, 2013) koristi se sledeći algoritam:

1) Obučava se NNLM sa listom najčešćih reči na izlaznom sloju. Ovaj korak potreban je kako bi se dobio početni model projekcionog prostora, koji je definisan matricom **R** (videti sliku 7).

2) Primenjuje se raščlanjivanje na osnovne komponente (*Principal Component Analysis* – PCA) radi redukcije dimenzionalnosti matrice **R**.

3) Izvršava se klasifikacija reči odozgo nadole (eng. *top-down clusterisation*) pomoću *K-means* algoritma korišćenjem kontinualne reprezentacije reči koje se ne nalaze na spisku najčešćih, a za koje je reprezentacija izvedena u prethodnom koraku.

4) Obučava se NNLM nad celim rečnikom, sa hijerarhijskom strukturom na izlazu.

Dakle, cilj prvog koraka navedenog algoritma jeste da se utvrde obeležja reči koja su neophodna za klasterizaciju. Autori opisanog istraživanja su u svom ranijem radu (Le et al, 2010) opisali metodu inicijalizacije reprezentacije reči (u kontinualnom prostoru). Ova metoda podrazumeva da se inicijalno sve reči projektuju u istu (na slučajan način određenu) tačku. Intuitivno, deluje smisleno sve reči u početku posmatrati kao veoma slične, kako bi do odvajanja klasa reči došlo ako za to zaista postoje razlozi u korpusu za obuku modela. Pokazuje se da

ovakvim pristupom algoritam klasterizacije brzo konvergira, tačnije nakon svega nekoliko iteracija.

U okviru matrice **R**, svakoj reči odgovara jedan vektor – vektor obeležja te reči. Nakon redukcije dimenzionalnosti matrice **R**, u drugom koraku algoritma obuke, ovi vektori su u opisanom istraživanju bili dimenzije 10, što je za posledicu imalo da je *K-means* algoritam za klasterizaciju mogao biti primenjen veoma efikasno.

Treći korak algoritma obuke podrazumeva podelu klase reči ukoliko ona sadrži više od W instanci, pri čemu je W empirijski određeno. Početna klasa se deli na $\lfloor\sqrt{W}\rfloor + 1$ potklasa. Eksperimenti su pokazali da je W od zanemarljivog uticaja na kvalitet ASR sistema koji se oslanja na SOUL NNLM i da je dubina klasifikacionog stabla $D = 3$ dovoljna. Naravno, ove podatke treba koristiti kao orijentacione, jer postoji mogućnost da za neke inflektivne jezike klasifikacija mora da bude znatno komplikovanija.

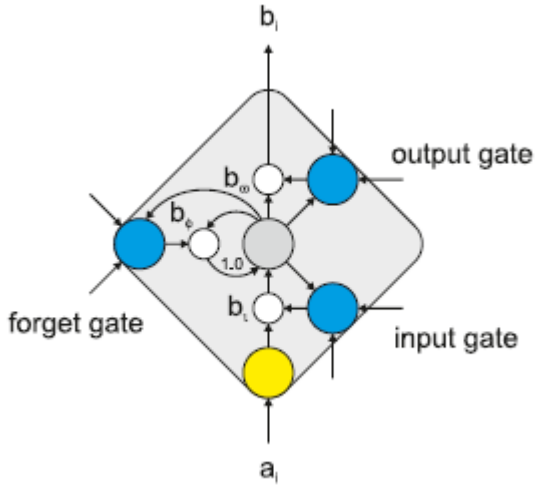
Način klasterizacije opisan u (Le et al, 2010) poređen je sa Braunovim algoritmom klasterizacije (Brown et al, 1992), o kome će biti nešto više reči u poglavlju 4, a koji se smatra *state-of-the-art* algoritmom (a zasniva se na maksimizaciji zajedničke informacije između susednih klasa) i klasterizacijom po učestanosti reči, koja je najjednostavnija. Pokazalo se da su predloženi i Braunov algoritam uporedivi po smanjenju perpleksnosti modela jezika.

U nekoliko eksperimenata koji su sprovedeni u okviru ovog istraživanja pokazalo se da SOUL struktura daje doprinos kvalitetu NNLM kada je u pitanju struktura mreže sa propagacijom unapred, ali istraživanja u odnosu na RNN strukturu još uvek nisu sprovedena, iako se zna da RNNLM sam po sebi prevazilazi mogućnosti FFNN modela.

LSTM NN modeli jezika

Motivacija za uvođenje LSTM (*Long Short-Term Memory*) jedinica neuronske mreže u obuku modela jezika vezana je za problem „nestajućeg gradijenta“, koji je posledica propagacije gradijenta greške unazad kroz vreme pri obuci pomoću *BPTT* algoritma. Iako postoje različiti načini na koje je ovaj problem rešavan, najčešće korišćenjem informacija višeg nivoa apstrakcije iz ulaznih podataka, uglavnom se rešavanjem postojećeg problema javlja problem prevelike kompleksnosti procesa obuke. Koncept LSTM u modelovanju jezika je predstavljen u (Sundermeyer et al, 2012), gde je ukratko opisan i proces kojim se došlo do nove strukture veštačkog neurona. Naime, pri propagaciji gradijenta greške vrednosti

gradijenta se skaliraju određenim faktorom. U praksi, ovaj faktor je gotovo uvek različite vrednosti od 1, zbog čega dolazi ili do nestajanja ili do eksplozije vrednosti gradijenata posle svega nekoliko koraka propagacije. Da bi se izbegao efekat skaliranja, neuronska jedinica je redizajnirana tako da faktor skaliranja bude fiksan i jednak 1. Neuronska jedinica dobijena ovim putem bila je značajno ograničena po pitanju mogućnosti „učenja“. Stoga je dalji razvoj LSTM jedinice podrazumevao je uvođenje nekoliko tzv. *gating* jedinica. Konačna struktura LSTM jedinice prikazana je na slici 8.

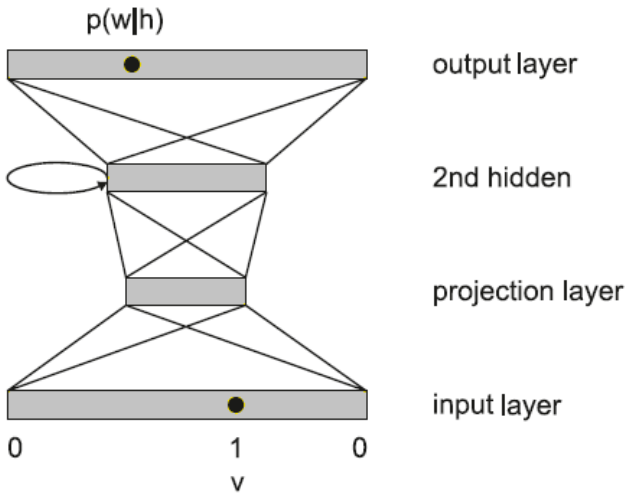


Slika 8. Struktura LSTM jedinice neuronske mreže

Kod standardnog neurona ulazna a_i i izlazna b_i aktivacija povezane su tangens hiperboličnom aktivacionom funkcijom $b_i = \tanh(a_i)$. Kod LSTM neurona, nakon što se na a_i primeni aktivaciona funkcija, rezultat se množi sa b_ι. Zatim se, preko rekurentne veze, na dobijenu vrednost dodaje unutrašnja aktivaciona vrednost iz prethodnog vremenskog koraka, koja je prethodno pomnožena sa b_ϕ. Na kraju, rezultat se skalira sa b_ω kako bi se dobila izlazna aktivaciona vrednost. Skalirajući faktori b_ι, b_ϕ, i b_ω imaju vrednosti veće od 0 a manje od 1, a kontrolisane su pomoću posebnih jedinica – ulazne kapije (eng. *input gate*), izlazne kapije (eng. *output gate*) i kapije zaborava (eng. *forget gate*). Kapija zaborava je od naročitog značaja, jer se uz pomoć nje kontroliše u kojoj meri je aktuelne podatke potrebno „pamtiti“. U (Sundermeyer et al, 2012) je prikazana osnovna struktura LSTM jedinice, ali je detaljni matematički opis LSTM NN predstavljen u ranijem istraživanju koje je opisano u (Hochreiter & Schmidhuber, 1997). Struktura mreže koja je korišćena za obuku modela jezika u (Sundermeyer et al, 2012) data je na slici 9.

Smatra se da se LSTM neuronska jedinica može posmatrati kao diferencijabilna verzija računarske memorije. Zato je često zovu i LSTM memorijska ćelija. Ovakva struktura rešava problem nestajućeg gradijenta kod RNN. Na slici 9 vidi se da su korišćena praktično dva skrivena sloja, a LSTM jedinice su postavljene u drugi, rekurentni sloj. U okviru istog istraživanja, između ostalog, rađeni su i eksperimenti kako bi se videlo da li dodatni slojevi mogu poboljšati konačni model jezika. Pokazano je da ni slojevi sa linearnim ni slojevi sa *sigmoid* aktivacionim funkcijama koji se dodaju na strukturu sa slike 9 ne donose dalja smanjenja perpleksnosti LM.

Rezultati po pitanju perpleksnosti pokazuju poboljšanje od oko 8% kada se koriste LSTM jedinice kod rekurentnog sloja. Ovo poboljšanje gotovo je konstantno kada se menja veličina rekurentnog skrivenog sloja. Značajno poboljšanje po pitanju *WER* (oko 0,5%) dobijeno je u odnosu na *Kneser-Ney* kvadrigram model, nakon interpolacije sa LSTM modelom, iako je ovaj drugi obučavan na daleko manje (oko 100 puta manje) podataka.



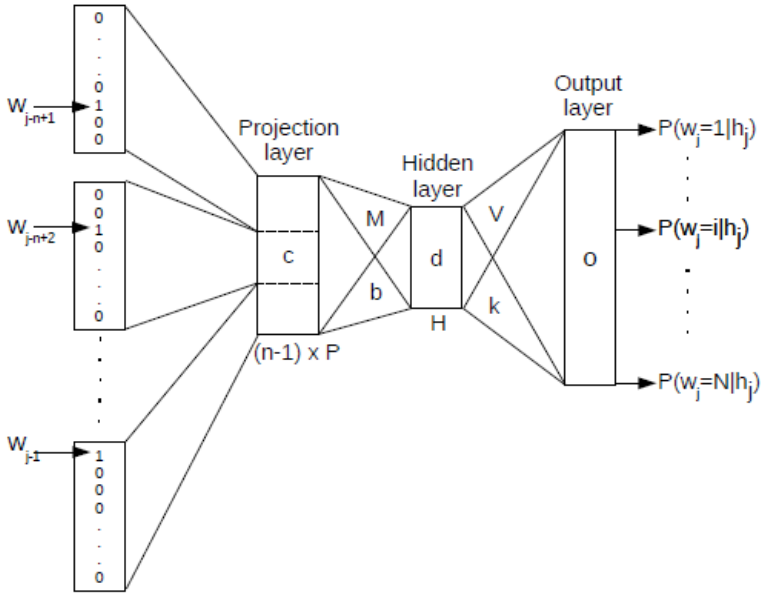
Slika 9. Struktura LSTM neuronske mreže

DNN modeli jezika

Istraživanje vezano za tzv. „duboke“ neuronske mreže (eng. *Deep Neural Networks*) rađeno je otprilike u isto vreme kada i LSTM istraživanje. Iako je LSTM eksperiment pokazao da dodavanje dodatnih slojeva na RNNLM sa LSTM

jedinicama na rekurentnom sloju ne dovodi do daljih poboljšanja po pitanju perpleksnosti modela, dodavanje dodatnih slojeva doprinosi kvalitetu NNLM bez rekurentne veze, što je pokazano u (Arisoy et al, 2012). Motivacija za ovo istraživanje proistekla je iz rezultata istraživanja vezanog za akustičko modelovanje, koje je pokazalo da su DNN uporedive sa GMM (*Gaussian Mixture Model*) po uspešnosti modelovanja.

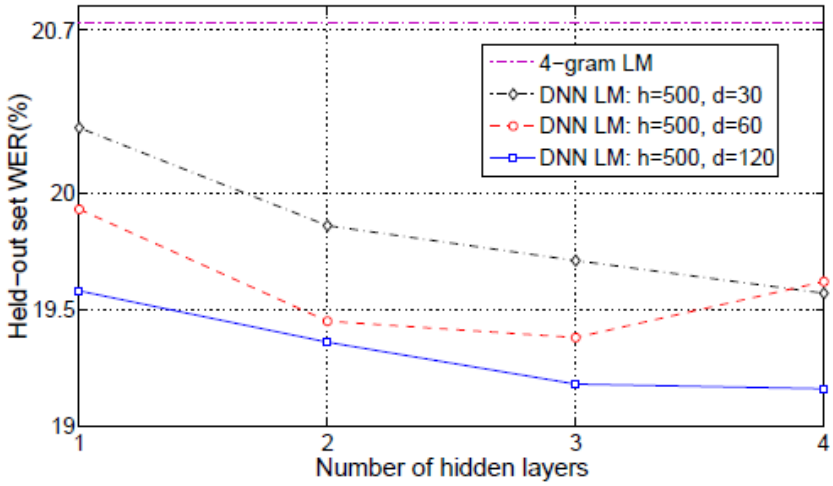
Eksperimenti su rađeni sa modelima baziranim na strukturi koja je prikazana na slici 10. Modeli su obučavani na korpusu koji se sastoji od oko 900 hiljada rečenica. Izlazni rečnik postavljen je na veličinu 10 hiljada. Svi modeli su interpolirani sa standardnim kvadrigram modelom radi ublažavanja. Eksperiment je obuhvatio obuku NN sa do četiri skrivena sloja. Dimenzije skrivenih slojeva su postavljene na $h = 500$, dok je dimenzionalnost projekcionog sloja d varirana. Rezultati su prikazani grafički na slici 11. Pokazano je, dakle, da dodavanje skrivenih slojeva na osnovnu strukturu doprinosi smanjenju *WER* za 0,2-0,4%. Što se tiče vrednosti perpleksnosti, rezultati su dati u tabeli 1 paralelno sa vrednostima za *WER* (Arisoy et al, 2012).



Slika 10. Struktura standardne NN sa jednim skrivenim slojem

Ono što je zanimljivo za primetiti jesu rezultati dobijeni za RNNLM. Za kreiranje ovih modela korišćen je *RNNLM toolkit*. S obzirom na to da ovaj alat pruža dodatna poboljšanja strukture modela radi efikasnosti izračunavanja, a

izlazni rečnik je jednake dimenzije sa ulaznim, što nije bio slučaj sa DNN modelima kreiranim u okviru ovog istraživanja, ove modele nije moguće porediti direktno po dobijenim *ppl* vrednostima. Stoga ostaje nejasno da li su RNNLM bolji od DNN LM, što je svakako zanimljiva tema za dalje istraživanje. Takođe, za DNN model se pokazalo da lako „upada“ u lokalno optimalno rešenje pri obuci, te se može pretpostaviti da bi prethodna obuka radi inicijalizacije DNN mogla značajno doprineti uspešnosti ove strukture u modelovanju jezika.



Slika 11. Rezultati poređenja NN sa različitim brojem skrivenih slojeva

modeli	*ppl*	*WER*(%)
kvadrigram LM	114,4	22,3
DNN LM: h = 500, d = 30 1 skriveni sloj (NNLM) 4 skrivena sloja	115,8 108,0	22,0 21,6
DNN LM: h = 500, d = 60 1 skriveni sloj (NNLM) 3 skrivena sloja	109,3 105,0	21,5 21,3
DNN LM: h = 500, d = 120 1 skriveni sloj (NNLM) 3 skrivena sloja	104,0 102,8	21,2 20,8
RNNLM (h = 200) RNNLM (h = 500)	99,8 83,5	- -

Tabela 1. Rezultati testiranja DNNLM po pitanju ppl i WER

2.3 Odnos *N*-gram modela i modela baziranih na neuronskim mrežama

Direktno poređenje *N*-gram i NN jezičkih modela predstavlja zadatak pri čijem rešavanju je lako prevideti određene činjenice, čime bi se mogla dati neopravdana inicijalna prednost jednom ili drugom tipu modela. Problem je najčešće uzeti u obzir veličine modela. Na veličinu NN modela može se uticati menjanjem broja neurona pojedinih slojeva mreže (u zavisnosti od strukture koja se koristi), čime se može vršiti grubo podešavanje veličine modela, dok se kod *N*-gram modela prvobitna veličina, koja zavisi od reda modela, može korigovati uklanjanjem određenog broja *N*-grama nekom od tehnika tzv. potkresivanja (eng. *pruning*), od kojih neke omogućavaju precizno podešavanje veličine modela. Zbog ovoga je pogodnije prilagoditi veličinu *N*-gram modela veličini NN modela, kada ih je potrebno porediti. S obzirom na to da kvalitet *N*-gram modela nakon potkresivanja zavisi od korišćene tehnike, poređenje ovakvog modela sa

odgovarajućim NN modelom ne može biti u potpunosti korektno. Ipak, vršena su istraživanja u vezi sa kombinovanjem pomenuta dva tipa modela, na osnovu kojih se došlo do nekih uopštenih zaključaka o tome u kojim se situacijama koji od modela bolje snalazi.

U jednom od ovih istraživanja (Oparin et al, 2012), krenulo se od pretpostavke da je bolji model onaj koji validnoj sekvenci reči dodeljuje veću verovatnoću. Intuitivno, procenjeno je da bi NN modeli trebali da budu u prednosti u situacijama kada *N*-gram modeli estimiraju verovatnoće pomoću *back-off* procedure. Za svaku reč u skupu za testiranje može se dobiti informacija o tome na kom *back-off* nivou je verovatnoća estimirana pomoću *N*-gram modela. Stoga je moguće posebno posmatrati kako se NNLM ponaša u odnosu na svaki *back-off* nivo posebno. Ono što treba da se posmatra po svakom od pomenutih nivoa jeste procenat slučajeva u kojima se NNLM ponaša bolje od *N*-gram modela, zatim *ppl* vrednost celokupnog teksta za testiranje, kao i suma apsolutnih razlika u verovatnoćama koje nekom „događaju“ (reči) dodeljuju dva poređena modela. Vrednosti *ppl* po nivoima moraju da zadovolje sledeće:

$$\exp\left(\sum_k \frac{N_k}{N} \ln(ppl_k)\right) = ppl\ . \tag{2.51}$$

U izrazu (2.51), *ppl* je vrednost perpleksnosti izračunata na celom korpusu za testiranje, ppl_k je perpleksnost na nivou k izračunata pomoću izraza (2.52), N_k je broj reči za koje je verovatnoća estimirana na *back-off* nivou k, a N je ukupan broj reči u korpusu za testiranje.

$$ppl_k = \exp\left(-\frac{1}{N_k}\sum_{w_k} \ln P(w_k|h_k)\right) \tag{2.52}$$

U izrazu (2.52), w_k su reči za koje su verovatnoće estimirane na osnovu *N*-grama na *k*-tom *back-off* nivou, a h_k su odgovarajući konteksti (istorije).

Dva su posebna slučaja na koje je očigledno potrebno obratiti pažnju. Prvi slučaj su *N*-grami koji se pojavljuju jednom u korpusu za obuku. Ublažavanje, naročito u slučaju *Kneser-Ney* postupka, može u praksi često dovesti do toga da jednom „viđeni“ *N*-grami dobiju verovatnoće veoma bliske nuli. Drugi slučaj predstavljaju *N*-grami kod kojih je broj pojavljivanja samih *N*-grama i odgovarajućih istorija (niz reči bez poslednje) jednak. Stoga su, u okviru analize performansi dva poređena LM, uzeti u obzir sledeći parametri:

1) broj pojavljivanja pojedinih *N*-grama ($w_1, w_2, \dots, w_N$);

2) broj pojavljivanja istorija za pojedine *N*-grame ($w_1, w_2, \dots, w_{N-1}$);

3) broj različitih reči w_N koje prate pojedine istorije.

Osnovni zaključci do kojih se došlo u okviru pomenutog istraživanja su sledeći:

1) NN modeli jezika (FFNN i RNN) daju veće verovatnoće za oko 50% *N*-grama koji se pojavljuju u korpusu za obuku u odnosu na standardni *N*-gram model. Ovaj procenat raste sa porastom nivoa *back-off* procedure.

2) Nema jasnih dokaza da NNLM daje bolje rezultate za *N*-grame koji su viđeni jednom u korpusu za obuku.

3) *N*-gram modeli daju bolje rezultate u slučajevima kada je broj pojavljivanja *N*-grama i odgovarajuće istorije jednak, odnosno kada se nakon određene sekvence u korpusu uvek pojavljuje ista reč.

4) Pri poređenju različitih struktura NN modela, RNN modeli dali su bolje rezultate od kvadrigram FFNN.

POGLAVLJE 3

RESURSI ZA OBUKU JEZIČKIH MODELA ZA SRPSKI JEZIK

Početak razvoja tehnologija koje su vezane za implementaciju veštačke inteligencije podrazumeva kreiranje baza podataka i drugih resursa na osnovu kojih se vrše obuke sistema. Tokom razvoja govornih tehnologija za srpski jezik, oformljeno je mnoštvo baza podataka i razvijeni su alati za obradu govornog signala, kao i tekstualnih sadržaja (Delić et al, 2013). Neki od tih resursa bili su od velikog značaja za razvoj jezičkih modela za srpski jezik i oni će biti ukratko predstavljeni u ovom poglavlju. Vremenski najzahtevniji deo ovog istraživanja predstavljalo je prikupljanje velike količine tekstualnih sadržaja na srpskom jeziku i priprema tih tekstova za obuku različitih tipova jezičkih modela. Proces prikupljanja tekstualnih sadržaja, rešavanje brojnih problema sa ciljem da se oformi relativno kvalitetan korpus, kao i alati koji su nastali u tom procesu, a koji će biti od značajne koristi pri budućim proširenjima nastalog korpusa, takođe će biti opisani u okviru ovog poglavlja.

3.1 Morfološki rečnik srpskog jezika

S obzirom na činjenicu da srpski jezik spada u grupu jezika sa kompleksnom morfologijom, za obuku kvalitetnih modela jezika potrebna je veoma velika količina podataka. Poređenja radi, za engleski, koji ne spada u grupu visoko inflektivnih jezika, postoji nekoliko tekstualnih korpusa koji sadrže preko 100 miliona reči (za američki engleski čak i preko 350 miliona) (URL1). Korpusi slične veličine postoje i za nemački, francuski i španski. Za većinu jezika, ipak, postojeći korpusi su značajno manji (10-30 miliona reči) (Mengusoglu & Deroo, 2001). Prikupljanje i priprema tekstualnih sadržaja predstavljaju veoma spor proces. U toku ovog istraživanja oformljen je tekstualni korpus za srpski jezik koji broji oko 23 miliona reči. Zbog toga je bilo neophodno kreirati adekvatne klasne modele. Pretpostavka je bila da je grupisanje reči na osnovu morfoloških kategorija pogodan način za delimično prevazilaženje problema male količine podataka za obuku. Način na koji su morfološke klase reči definisane detaljno će biti opisan u poglavlju 4, dok će u poglavlju 5 biti pokazano da grupisanje na osnovu morfoloških informacija daje bolje rezultate od uobičajenog načina izvođenja klasa pomoću bigram statistika.

Za kreiranje morfoloških klasa, neophodno je posedovati odgovarajuće resurse. Jedan od njih je morfološki rečnik, koji je nastao u toku ranijeg razvoja

govornih tehnologija za srpski jezik, pre svega sinteze govora na osnovu teksta (Sečujski, 2002). Ovaj rečnik se sastoji od preko 4 miliona unosa, ali se kontinuirano radi na njegovom proširenju. Kratak izvod iz ovog rečnika prikazan je na slici 12.

```
2 mesti meti [/0] 1 0 3 1 0 0 0 1
0 metak metka ["0] 0 1 3 0 0 0 0 0
0 metak metka ["0] 0 1 1 0 0 0 0 0
0 metak metke ["0] 0 1 3 1 0 0 0 0
0 metak metkom ["0] 0 1 5 0 0 0 0 0
0 metak metku ["0] 0 1 4 0 0 0 0 0
0 metak metku ["0] 0 1 2 0 0 0 0 0
0 metla metla [\0] 1 1 0 0 0 0 0 0
0 metla metlama [\00] 1 1 5 1 0 0 0 0
0 metla metlama [\00] 1 1 2 1 0 0 0 0
0 metla metle [\0] 1 1 4 1 0 0 0 0
0 metla metle [\0] 1 1 3 1 0 0 0 0
0 metla metle [\0] 1 1 0 1 0 0 0 0
0 metla metle [\0] 1 1 1 0 0 0 0 0
0 metla metli [\0] 1 1 1 1 0 0 0 0
0 metla metli [\0] 1 1 2 0 0 0 0 0
```

Slika 12. Izvod iz morfološkog rečnika srpskog jezika

Svaki od unosa rečnika počinje kodom za vrstu reči (0 – imenica, 1 – zamenica, 2 – glagol, 3 – pridev, 4 – broj, 5 – prilog, 6 – predlog, 7 – veznik, 8 – partikula, 9 – uzvik), nakon čega slede osnovni oblik reči (lema), izvedeni (pojavni) oblik reči i informacija o akcentu (u ugaonim zagradama po jedan simbol za svaki slog, pri čemu 0 označava nenaglašen slog, dok se za četiri vrste akcenta u srpskom jeziku koriste simboli „"“, „^“, „\“ i „/“ za kratkosilazni, dugosilazni, kratkouzlazni i dugouzlazni, respektivno) nakon čega slede morfološke informacije. Morfološke kategorije i druga obeležja predstavljena su pomoću osam brojeva, čije značenje varira u zavisnosti od vrste reči. Za imenice, na primer, relevantne informacije sadržane su u prva četiri broja, a značenja su im rod, vrsta, padež i broj, respektivno. Kod predloga se, s druge strane, koriste prvih pet brojeva i to kao binarne kategorije, a svaki od njih daje odgovor na pitanje da li je u pitanju predlog koji ide uz određeni padež imenice (genitiv, dativ, akuzativ, instrumental, nominativ).

Kao što se vidi na slici 12, za jedan pojavni oblik reči može postojati više unosa u rečniku. Pri klasifikaciji reči na osnovu morfoloških informacija, potrebno je za svaku odrediti koji unos u rečniku joj odgovara, odnosno neophodno je izvršiti morfološku anotaciju teksta. O ovom procesu će viti više reči u odeljku 3.2.2.

3.2 Prikupljanje i pretprocesiranje tekstualnih sadržaja

Formiranje tekstualnog korpusa za potrebe obuke jezičkih modela predstavljalo je dugotrajan proces sakupljanja tekstova i adaptaciju istih. Tekstovi su klasifikovani na osnovu funkcionalnih stilova, te su formirani odvojeni korpusi za novinarski (publicistički), literarni (književno-umetnički), naučni, naučno-popularni, administrativni (birokratski) i govorni (razgovorni) stil.

Za formiranje korpusa za novinarski stil prikupljena je velika količina novinskih članaka, različitih tematika. Za literarni stil izdvojeni su tekstualni sadržaji mnoštva romana i kratkih priča. Za naučni stil poslužili su sadržaji diplomskih radova, doktorskih disertacija, naučnih i stručnih radova i drugih publikacija iz različitih oblasti. Za naučno-popularni stil korišćeni su tekstovi iz naučno-popularnih časopisa. Za administrativni stil izdvojen je sadržaj Ustava Republike Srbije, tekstovi zakona, ugovora, molbi, žalbi, sudskih rešenja itd. Za govorni stil prikupljeni su titlovi za filmove.

Nakon prikupljanja tekstova, zaključeno je da je korpus za naučno-popularni stil isuviše mali da bi bio od koristi u modelovanju jezika, te je ovaj korpus pripojen korpusu za naučni stil. Korpus za razgovorni stil je takođe premali za jezičko modelovanje, te nije korišćen u daljem istraživanju, iako je izvršena priprema prikupljenog sadržaja, sa ciljem da se ovaj korpus u budućnosti nadograđuje.

Važno je napomenuti da bi, u idealnom slučaju, bilo od koristi podeliti korpuse za obuku po temama, što bi omogućilo kreiranje posebnih jezičkih modela za svaku od tema, a zatim interpolaciju određenog broja modela sa različitim težinskim koeficijentima. Na ovaj način bi se rezultujući model najbolje prilagodio određenom zadatku. Ipak, za ovakvu podelu korpusa potrebno je imati mnogo više podataka. Pored tema, od značaja je i podatak o vremenu u kom je tekst nastao. Ovo je naročito važno za literarni stil, gde se rečnik, pa i konstrukcija rečenica razlikuje od autora do autora, ali posebno ako su stvarali u različitim vremenskim periodima. Na žalost, i ovakva podela je smislena samo ako su na raspolaganju ogromne količine podataka.

Priprema prikupljenih tekstova za modelovanje jezika obuhvatala je usaglašavanje mnogih konvencija, čišćenje sadržaja od nerelevantnih elemenata i slično. Ovaj proces odvijao se u više iteracija i podrazumevao je manuelne, poluautomatske i automatske korekcije tekstova. Manuelna korekcija podrazumevala je pregledanje prikupljenih sadržaja i izbacivanje delova koji bi narušavali statistike, kao što su, na primer, tabelarni prikazi sportskih rezultata, koji su bili sastavni deo tekstova sportskih vesti, matematički izrazi, koji se često

pojavljuju u naučnim tekstovima, sadržaji koji su u procesu prikupljanja zbog određenog inicijalnog formata izgubili smisao itd. U toku pregledanja sadržaja uočavani su i detalji koji su mogli biti korigovani korišćenjem regularnih izraza, što predstavlja poluautomatsku korekciju (određeni rezidualni sadržaji parsiranih xml dokumenata i slično). Određene korekcije bilo je, ipak, najjednostavnije rešiti automatski, implementacijom alata za obradu tekstualnih korpusa, u okviru kojeg su realizovane različite funkcionalnosti, poput npr. konverzije pisma (s obzirom na to da su neki od tekstova bili na ćiriličnom, a neki na latiničnom). Alat koji je za ove potrebe razvijen u okviru ovog istraživanja zove se *Txtproc*, a neke od njegovih funkcionalnosti će biti detaljnije opisane u narednom odeljku.

3.2.1 ALAT *TXTPROC*

Alat *Txtproc* je inicijalno kreiran za potrebe pripreme tekstualnih sadržaja za obuku modela jezika. Međutim, u toku razvoja jezičkih modela i njihove primene u različitim tehnologijama, ovom alatu su dodavane nove funkcionalnosti, te su one vremenom grupisane u više odvojenih izvršnih programa. Izvršni programi i njihove funkcionalnosti koje su od značaja za ovo istraživanje će biti navedene u nastavku.

TxtprocSplitAndMergeFiles

Za potrebe poluautomatskog uređivanja tekstova korišćenjem regularnih izraza bilo je potrebno kreirati relativno male dokumente, kako bi se optimizovala brzina rada u editoru *Notepad++*. Takođe, bilo je potrebno omogućiti izdvajanje željenog procenta sadržaja za evaluaciju modela. Stoga je, u okviru ovog izvršnog programa, implementirano nekoliko funkcionalnosti.

1) Podela dokumenta na željeni broj delova. Pri tome je omogućena podela na dva načina. Prvi podrazumeva da se u svaki od N željenih rezultujućih dokumenata snimi $1/N$ uzastopnih rečenica iz originalnog dokumenta. Drugi podrazumeva snimanje svake N-te rečenice u određeni rezultujući dokument. Potreba za implementacijom ova dva načina podele dokumenta vezana je za utvrđivanje toga koliko su rezultujući dokumenti reprezentativni podskupovi sadržaja originalnog dokumenta, s obzirom na to da u originalnim dokumentima, kada su u pitanju korpusi za pojedine funkcionalne stilove, susedne rečenice često potiču iz istog izvora, odnosno pripadaju istoj tematici.

2) Spajanje dokumenata u jedan.

3) Izdvajanje određenog procenta podataka radi formiranja test skupa. Ovo zapravo predstavlja podelu dokumenta na dva dela različitih veličina.

TxtprocProcessingForASR

Eksperimenti vezani za prepoznavanje govora podrazumevali su različite manipulacije tekstualnim sadržajima. Implementirane funkcionalnosti, koje su vezane su za jezičke modele i njihovu primenu u ASR, biće opisane u nastavku.

1) Izdvajanje teksta iz transkripcija za govorne baze. Naime, za potrebe prepoznavanja govora, prikupljena je velika količina audio-knjiga, snimaka radio-emisija i sličnih sadržaja, za koje su manuelno kreirane transkripcije sa posebnim oznakama za oštećenja na akustičkom, leksičkom i lingvističkom nivou (Suzić et al, 2014). Ova funkcionalnost omogućila je izdvajanje sadržaja koji je upotrebljiv za jezičko modelovanje.

2) Konverzija određenih simbola radi očuvanja sadržaja pri promeni načina kodovanja teksta (UTF-8 u ANSI i obrnuto).

3) Čišćenje teksta uklanjanjem nepoželjnih simbola, pri čemu se skup dozvoljenih simbola može proizvoljno odabrati.

4) Proširenje rečnika prepoznavača govora na osnovu tekstualnog korpusa. Ova funkcionalnost je od velikog značaja pri prilagođavanju sistema za prepoznavanje govora određenom domenu primene.

TxtprocLMCorpusProcessing

U okviru ovog izvršnog programa sadržane su funkcionalnosti koje su od značaja za različite eksperimente koji uključuju jezičke modele, a koji nisu implementirani u okviru *SRILM* alata. O svim ovim eksperimentima će biti više reči u poglavlju 5.

1) Promena redosleda reči u okviru rečenice, ili na nivou celokupnog tekstualnog dokumenta, na slučajan način. Ako je u pitanju premeštanje na nivou rečenice, postoji mogućnost premeštaja reči tako da se dobiju rečenice sa obrnutim redosledom reči.

2) Izdvajanje reči u odvojene linije teksta, uz korekcije radi dobijanja validnih rečenica od kojih svaka sadrži po jednu reč. Ova funkcionalnost bila je neophodna pri morfološkoj anotaciji u situacijama kada je bilo potrebno zanemariti kontekst.

3) Evaluacija segmentacije teksta na rečenice. Naime, *SRILM* pruža mogućnost primene jezičkog modela u segmentaciji teksta na rečenice, ali ne i procenu tačnosti segmentacije. Ova funkcionalnost, naravno, podrazumeva postojanje referentnog korpusa u kome je podela teksta na rečenice izvršena manuelno.

4) Zamena određenih reči u tekstu odgovarajućim klasama, na osnovu datog skupa relacija klasa-reči. Na ovaj način mogu se testirati detalji grupisanja reči, odnosno mogu se dijagnostikovati problemi u postupku grupisanja. U opštem slučaju, ova funkcionalnost može da služi za redukciju skupa klasa reči.

5) Formiranje skupa relacija klasa-reči na osnovu korpusa koji sadrži reči i korpusa koji umesto reči sadrži odgovarajuća imena klasa.

6) Dodavanje *XML* zaglavlja i konverzija određenih simbola u okviru pripreme teksta za snimanje u *XML* formatu. Ova funkcionalnost bila je neophodna radi usklađivanja ulaznih podataka sa očekivanim formatom ulaza za alat *anTagger*, koji će biti opisan u odeljku 3.2.2.

7) Konverzija ćiriličnog pisma u latinično i obrnuto (transliteracija).

8) Konverzija simbola za slova sa dijakriticima, koji su deo posebne konvencije, na koju se nailazilo u pojedinim dokumentima (*cc→ć*, *ss→š*, *zz→ž*, *ch→č*).

9) Izdvajanje teksta iz titlova za filmove. Ovo podrazumeva nekoliko koraka čišćenja početnog sadržaja. Pre svega, uklanjaju se informacije o vremenskim trenucima u kojima bi trebalo da budu prikazane tekstualne poruke u toku video snimka. Nakon što je izdvojen koristan tekst, vrši se provera i eliminišu se sve rečenice za koje se detektuje da nisu napisane na srpskom jeziku. Pojava da se u titlovima sporadično pojavljuju i neprevedeni delovi teksta uočena je tokom manuelnog pregledanja i korekcije sadržaja ovog tipa. Implementirana je detekcija ruskog i engleskog jezika, na osnovu višestrukog pojavljivanja simbola koji nisu deo latiničnog ili ćiriličnog pisma srpskog jezika u okviru jedne rečenice. Takođe, implementirana je detekcija hrvatskog jezika, uporednim proverama rečnika za srpski i hrvatski. Naposletku, vršeno je spajanje susednih segmenata teksta, a zatim podela dobijenog sadržaja na rečenice.

10) Tretman slova *đ*, koje se, u inače dijakritizovanim tekstovima povremeno greškom pojavljivalo u obliku *dj*. Zamenom *dj→đ* u

tekstovima bi došlo do mnogih grešaka, a većina njih javljala bi se kod složenica sa prefiksima poput predloga *od*, *pod*, *nad*, *pred*. Primeri ovakvih reči su *odjednom*, *podjednako*, *nadjačati*, *predjelo*. Zbog toga je tretman *đ* podrazumevao konverziju *dj→đ* kod reči koje ne počinju predlogom koji se završava slovom *d*, izuzev reči koje predstavljaju izuzetke od ovog pravila, a koje su snimljene u poseban dokument. Neki od primera ovakvih izuzetaka su *Nađ*, *nađen*, *pođemo*, *poređana*, *pređašnji*. Postoje, naravno, reči koje ne počinju pomenutim predlozima, ali konverzija *dj→đ* ipak nije potrebna, iako ih nije mnogo i retko se pojavljuju u tekstovima. Ovakve reči su takođe snimljene u poseban dokument, a neki primeri su prezimena *Zdjelar*, *Serdjukov*, *Kirdjapkin*. Skraćenica za *disk džokej* – *DJ* u svim pojavnim oblicima detektovana je i tretirana kao poseban slučaj. Na ovaj način se uz pomoć par pravila i nekoliko relativno kratkih spiskova izuzetaka mogla izvršiti konverzija *dj→đ*, gde je to bilo potrebno, bez potrebe da se vrši pretraga po rečniku, koja bi bila značajno sporija.

TxtprocFindUnique

Ovaj izvršni program služi za analizu sadržaja tekstualnih korpusa, a implementirane funkcionalnosti su korišćene i za potrebe eksperimenata vezanih za automatsko prepoznavanje govora.

1) Izdvajanje liste različitih reči koje se pojavljuju u dokumentu. Ovu funkcionalnost ima i *SRILM*, ali samo za dokumente u ANSI formatu.

2) Izdvajanje liste različitih grafema koji se pojavljuju u dokumentu. Ova funkcionalnost bila je od pomoći pri manuelnom čišćenju tekstualnih sadržaja.

3) Detekcija duplikata, odnosno rečenica koje se pojavljuju dva (ili više) puta. Naravno, za validaciju rezultata detekcije duplikata neophodna je manuelna kontrola.

TxtprocExtractStatistics

Kao i *TxtProcFindUnique*, i ovaj izvršni program je kreiran za potrebe analize sadržaja tekstualnih korpusa.

1) Izdvajanje informacija o učestanosti pojavljivanja pojedinih grafema.

2) Izdvajanje informacija o učestanosti pojavljivanja pojedinih reči.

3) Izdvajanje željenog broja reči koje se pojavljuju određeni broj puta. Ova funkcionalnost korišćenja je za proširenja rečnika sistema za automatsko prepoznavanje govora, a u vezi je sa funkcionalnošću 4 programa *TxtprocProcessingForASR*.

3.2.2 ALAT *ANTAGGER*

U okviru ranijih istraživanja, prvenstveno za potrebe sinteze govora na osnovu teksta na srpskom jeziku, razvijani su sistemi za automatsku morfološku anotaciju teksta. Jedan od ovih sistema je baziran na skrivenim Markovljevim modelima (Sečujski, 2009). Međutim, ovakav sistem je veoma ograničen po pitanju kompleksnosti pravila koja može da modeluje na osnovu podataka za obuku. Drugi sistem je baziran na transformacionim pravilima (Delić et al, 2009). Taj sistem, s druge strane, ne pruža mogućnost da se za neko pravilo odredi sa kojom pouzdanošću ga treba primeniti. Treći, ekspertski sistem, koji se pokazao kao najtačniji, poseduje pozitivne osobine prva dva i on se koristi u *AlfaNum* sistemu za sintezu govora na osnovu teksta (Sečujski et al, 2002). Ovaj alat takođe je poslužio kao osnova za kreiranje alata *anTagger*, koji je u okviru ovog istraživanja razvijen za potrebe pripreme korpusa za obuku različitih tipova modela, nakon što je tekstualni sadržaj prethodno očišćen i adaptiram korišćenjem funkcionalnosti alata *Txtproc*. U nastavku će biti opisane funkcionalnosti alata *anTagger*.

1) Segmentacija teksta na rečenice. Ova funkcionalnost se zapravo koristi u okviru pretprocesiranja tekstualnog sadržaja pomoću alata *Txtproc*, kao jedan od međukoraka, u okviru koga se prvo vrši odvajanje znakova interpunkcije od reči, čime se dobija niz elemenata teksta. Segmentacija na rečenice se dalje vrši analizom konteksta, odnosno susednih elemenata teksta, uz pomoć ručno implementiranih pravila.

2) Uklanjanje interpunkcijskih znakova. Ova funkcionalnost se koristi u slučaju da se korpus priprema za, na primer, obuku modela jezika koji će se koristiti u okviru sistema za automatsko prepoznavanje govora. U tom slučaju, jednom kada je tekst podeljen na rečenice, potrebno je zadržati samo sekvence reči. Ako se, pak, korpus priprema za obuku modela za potrebe korekcije grešaka u tekstovima, ova funkcionalnost nije od koristi.

3) Lematizacija. Ova funkcionalnost podrazumeva zamenu svakog pojavnog oblika reči odgovarajućim osnovnim oblikom – lemom. Podrazumeva se da ulazni dokument sadrži i znake interpunkcije, kako bi se mogla izvršiti adekvatna analiza konteksta.

4) Konverzija brojeva u reči. Ova procedura zahteva analizu konteksta, kao što to zahteva i lematizacija.

5) Konverzija reči u morfološke klase. Ovo zapravo predstavlja kreiranje korpusa za obuku klasnih modela jezika baziranim na morfološkim klasama reči. U poglavlju 4 će biti detaljno opisan način na koji su definisane morfološke klase reči.

6) Kreiranje korpusa za obuku hibridnih modela jezika. Ova funkcionalnost podrazumeva lematizaciju, konverziju reči u morfološke klase, a zatim kreiranje posebne vrste korpusa čiji su elementi uređene trojke ($reč, lema, morf.klasa$). O ovim modelima će takođe biti reči u poglavlju 4.

3.3 Sadržaj tekstualnih korpusa

Kao što je u uvodnom delu ovog poglavlja napomenuto, tekstualni korpus koji je prikupljen u okviru ovog istraživanja sastoji iz delova od kojih je svaki namenjen modelovanju određenog funkcionalnog stila. Pri tome, korpusi koji odgovaraju naučnom i naučno-popularnom stilu su, zbog malog obima i jednog i drugog, kao i sličnosti samih stilova pisanja, u eksperimentima bili tretirani kao jedan korpus, dok je korpus koji odgovara razgovornom stilu isuviše mali, te nije korišćen u eksperimentima. Svaki od delova korpusa se čuva u različitim oblicima, dobijenim nakon svake od ključnih etapa obrade. Tako se čuvaju korpusi nakon inicijalne obrade (čišćenje, unifikacija pisma itd.), nakon segmentacije teksta na rečenice, nakon konverzije brojeva u reči i nakon uklanjanja znakova interpunkcije. Verzija korpusa koja se koristi u određenom eksperimentu zavisi od namene modela jezika koji se kreira. Pored standardnih korpusa, koji sadrže pojavne oblike reči, čuvaju se i korpusi koji umesto pojavnih oblika reči sadrže odgovarajuće leme ili morfološke klase. Takođe, kreirani su i posebni korpusi za tzv. hibridne modele jezika, o kojima će biti više reči u narednom poglavlju, a primeri rečenica za svaki od pomenutih tipova korpusa dati su u prilogu 3. U tabeli 2 su prikazani detalji vezani za sadržaj pojedinih delova korpusa. S obzirom na to da su naučni i naučno-popularni korpusi u eksperimentima korišćeni kao jedan korpus kojim se modeluje naučni stil, uz podatke iz tabele 2 važni su i podaci o korpusu koji se dobija pomenutim združivanjem. Ovaj korpus sadrži 40.691 rečenica, 863.237 reči, 62.714 pojavnih oblika reči, 27.972 osnovna oblika reči – leme i 850 morfoloških klasa.

Na osnovu podataka iz tabele 2 može se doći do zanimljivih zaključaka o razlikama između funkcionalnih stilova, a samim tim i o važnosti vođenja računa o domenu primene modela jezika koji se kreiraju. Za početak, postoje značajne

razlike u prosečnim dužinama rečenica. Naime, dok su ove dužine između 20 i 25, odnosno približno jednake za novinarski (23,59), naučni (21,85), naučno-popularni (20,37) i administrativni (25,45) stil, prosečan broj reči u rečenici kod literarnog stila je 13,04, dok je kod govornog stila samo 3,4. S obzirom na velike razlike u obimima pojedinih delova korpusa, oni se ne mogu se direktno porediti po odnosima broja reči i pojavnih oblika reči (veličini rečnika), lema i morfoloških klasa koje se u njima pojavljuju, ali su neki od detalja ipak prilično očigledni. Na primer, vidi se da se najviše različitih morfoloških klasa pojavljuje u korpusu za literarni stil, iako je obim ovog korpusa više nego četiri puta manji od obima korpusa za novinarski stil, što oslikava raznolikost u rečeničnim strukturama koje karakteriše literarni stil. S druge strane, može se primetiti da iako je korpus za administrativni stil čak oko tri puta obimniji od korpusa za govorni stil, kod korpusa za administrativni stil se pojavljuje značajno manji broj morfoloških klasa, što govori o relativnoj jednoličnosti rečeničnih konstrukcija koja karakteriše administrativni stil. To navodi na zaključak da je najmanje podataka potrebno za obuku modela jezika koji bi se koristio u domenu u kom je u upotrebi administrativni stil, a najviše podataka za modelovanje literarnog stila. Ovi zaključci potvrđeni su i eksperimentima o kojima će biti više reči u poglavlju 5. Još jedan zanimljiv detalj jeste porast broja morfoloških klasa koje se pojavljuju u korpusu sa oko 770 na 850 pri združivanju korpusa za naučni i naučno-popularni stil. Po svim ostalim podacima može se primetiti da su ovi stilovi vrlo slični, ali na osnovu ovog podatka, evidentno je da ih ima smisla tretirati odvojeno kada se za to stvore uslovi, odnosno, kada se korpusi u dovoljnoj meri prošire.

korpusi	rečenica	reči	pojavnih oblika	lema	morf. klasa
novinarski	736.666	17.383.052	312.737	151.203	1.053
literarni	303.026	3.952.277	183.576	71.698	1.063
naučni	23.122	505.185	48.041	22.287	777
naučno-popularni	17.569	358.052	30.104	13.051	773
administrativni	14.888	378.834	18.589	7.237	621
govorni	37.690	128.235	14.875	8.403	690
ukupno	1.132.961	22.705.635	420.752	193.908	1117

Tabela 2. Sadržaj tekstualnih korpusa za srpski jezik

POGLAVLJE 4

MORFOLOŠKE KLASE REČI

S obzirom na činjenicu da je prikupljena količina podataka za obuku jezičkih modela za srpski jezik relativno mala u odnosu na korpuse koji se koriste za neke druge, pri tome manje kompleksne jezike u morfološkom smislu, potrebno je redukovati veličinu rečnika grupisanjem reči, tako da se omogući adekvatnija (indirektna) estimacija verovatnoća sekvenci reči pomoću dobijenih klasa.

Najpoznatiji algoritam za automatsku klasterizaciju reči je Braunov algoritam, koji se, kao što je pomenuto u odeljku 2.2.4, oslanja na bigram statistike tekstualnog korpusa. Prvobitno definisani algoritam bio je kompleksnosti $O(V^3)$, gde je V veličina rečnika. Ovaj algoritam podrazumeva da se inicijalno svaka reč tretira kao posebna klasa. Iterativnom primenom nekog od takozvanih „pohlepnih" (eng. *greedy*) algoritama, klase s spajaju dok se ne dostigne unapred zadati broj klasa. Zbog velike kompleksnosti ovog algoritma, uobičajeno je da se u procesu spajanja klasa koriste bigram statistike. Ipak, i u tom slučaju, originalni algoritam je vremenski veoma zahtevan. Optimizovana verzija Braunovog algoritma (Brown et al, 1992), čija je kompleksnost $O(VC^2)$, podrazumeva postavljanje parametra C, odnosno broja klasa, od kojih inicijalno svaka sadrži jednu od C reči koje se najčešće pojavljuju u korpusu. Potom se, u svakoj od iteracija, jedna reč dodeljuje nekoj od klasa.

Braunov algoritam pokazao se pogodnim za izdvajanje sintaksnih i semantičkih klasa reči. Očigledan problem je, međutim, to što se algoritmu mora zadati broj klasa reči, zbog čega je često potrebno primenjivati algoritam više puta, kako bi se približno odredio adekvatan broj klasa reči za dati tekstualni korpus. Pred toga, bigram statistike postavljaju ograničenja po pitanju određivanja sličnosti reči i logično je pretpostaviti da bi se uzimanjem u obzir dužeg konteksta, adekvatnije izvršila klasterizacija reči. Na kraju, razdvajanje najfrekventnijih reči u odvojene klase pri inicijalizaciji algoritma ne predstavlja postupak zasnovan na empirijskom znanju, već prosto predstavlja jedan od mnogih načina na koje bi se ovakav algoritam mogao inicijalizovati, a koji je izabran zbog jednostavnosti. Ipak, za engleski jezik, za koji je rađeno najviše istraživanja na ovu temu, još pre nekoliko decenija, ovaj algoritam dao je sasvim zadovoljavajuće rezultate.

S obzirom na postojanje veoma kvalitetnih jezičkih resursa za sprski jezik, koji bi se, između ostalog, mogli iskoristiti i za grupisanje reči, kao i zbog same činjenice da se može očekivati da Braunov algoritam za srpski, zbog morfološke

kompleksnosti jezika, ne pruži dovoljno dobre rezultate, došlo se na ideju da se definišu klase reči na osnovu morfoloških informacija. Naime, pomoću alata *anTagger* i na osnovu podataka koji su sadržani u morfološkom rečniku srpskog jezika, omogućeno je grupisanje reči na osnovu morfologije, kao i automatska konverzija tekstualnih korpusa u korpuse koji sadrže identifikatore morfoloških klasa i kao takvi služe za obuku klasnih jezičkih modela za srpski jezik.

4.1 Definisanje morfoloških klasa

Za potrebe ovog istraživanja, definisane su ukupno 1.124 morfološke klase reči. Klase su definisane na osnovu informacija koje su sadržane u morfološkom rečniku, a koje su predstavljene skupom oznaka – tagova (eng. *tag*). Pri definisanju morfoloških klasa, iskorišćena je većina tagova koji postoje u rečniku, mada su pojedini tagovi grupisani, na osnovu empirijskog znanja i eksperimentalnih rezultata (ukoliko se, dakle, pokazalo da se ništa ne postiže njihovim razdvajanjem, kada je u pitanju kvalitet jezičkih modela). Ovde je važno napomenuti da jezičke podele, koje su opisane skupom tagova u morfološkom rečniku, odstupaju u izvesnoj meri od standardnih jezičkih podela, jer su tagovi definisani imajući u vidu praktičnu upotrebu, odnosno specifičnosti primene morfoloških informacija u sistemima zasnovanim na govornim i jezičkim tehnologijama. U nastavku će biti detaljnije opisano kako su tretirane pojedine vrste reči, odnosno koje morfološke kategorije su za pojedine vrste reči uzete u obzir prilikom definisanja klasa.

Imenice. Relevantne morfološke kategorije za klasifikaciju ove vrste reči su padež, broj i tip. Kao obeležje koristi se i rod imenice, a tipovi imenica koji su uzeti u obzir su vlastite (posebni skupovi klasa definisani su za imena, prezimena, nazive organizacija i toponime), zajedničke, zbirne, gradivne i apstraktne.

Zamenice. Morfološka obeležja koja su korišćena uključuju padež, broj, rod, lice, a od ostalih obeležja koristi se još i tip. Naravno, nisu sva obeležja primenjiva na sve tipove zamenica. Na primer, lice je primenljivo jedino kod ličnih zamenica. Empirijski je, takođe, utvrđeno da je neke grupe, ili čak pojedinačne zamenice pogodno smestiti u posebne klase, što se uglavnom odnosi na prisvojne i upitno-odnosne zamenice (primeri zamenica koje su definisane kao zasebne klase su *sebe*, *sebi*, *sobom*, *čija*, *koja*, *kakva*, *kolika*).

Glagoli. Obeležja koja se koriste (ukoliko su primenjiva) su broj, rod i lice, kao i to da li je glagol prelazni ili neprelazni i da li spada u nepovratne, nekad povratne ili povratne glagole. Neki pojedinačni glagoli čine posebne klase (npr. glagol *nemoj*) zbog specifičnog morfološkog statusa.

Pridevi. Korišćena obeležja su stepen poređenja prideva, padež, broj i rod. Svi nepromenljivi pridevi čine zasebnu klasu. Pridev *nalik* je zbog specifičnog morfološkog statusa jedini pridev koji je izdvojen u zasebnu klasu.

Brojevi. Korišćena morfološka obeležja obuhvataju padež, broj i rod. Kod brojeva postoje mnogi izuzeci u odnosu na standardne gramatičke podele kada je reč o formiranju morfoloških klasa. Na primer broj *jedan* tretira se odvojeno (u zavisnosti od padeža, roda i broja može imati jedan od 18 oblika i svaki od njih čini po jednu morfološku klasu). Nepromenjivi brojevi čine jednu morfološku klasu. Formirana je i klasa „ostalo" za veoma retke slučajeve (npr. *četirima*).

Prilozi, *veznici*, *rečce*. Klase su formirane empirijski. Kod većine reči ovih vrsta svaka reč predstavlja klasu za sebe.

Predlozi. Klasifikacija se vrši prema padežu ili padežima imenice sa kojom predlog formira predloško-padežnu konstrukciju.

Uzvici. Svi uzvici čine jednu morfološku klasu.

4.2 Jezički modeli na bazi morfoloških klasa

Morfološkom anotacijom originalnog tekstualnog korpusa i klasifikacijom reči na osnovu dobijenih informacija dobija se korpus koji, umesto reči, sadrži identifikatore prethodno definisanih morfoloških klasa reči. Ovakav korpus se, na identičan način kao i originalni korpus, može koristiti za obučavanje jezičkih modela, pomoću alata kao što su *SRILM* i *RNNLM*. Modeli koji se dobijaju na ovaj način nisu kompletni bez informacija o verovatnoćama pripadanja pojedinih reči određenim klasama. Drugim rečima, verovatnoća pojavljivanja reči w_n koja pripada klasi c_n, nakon sekvence $w_1, \dots, w_{n-1}$ kojoj odgovara sekvenca klasa $c_1, \dots, c_{n-1}$, izračunava se pomoću izraza (4.1):

$$P(w_n|w_1 \dots w_{n-1}) = P(w_n|c_n)P(c_n|c_1 \dots c_{n-1}). \quad (4.1)$$

O ovome je potrebno voditi računa i pri evaluaciji klasnih modela jezika, odnosno pri računanju perpleksnosti na skupu podataka za testiranje. Klasni jezički model, dakle, ima $C^n - 1 + V - C$ nezavisnih parametara, gde je V veličina originalnog rečnika, a C broj klasa. U zavisnosti od primene modela jezika, odnosno u zavisnosti od memorijskih i drugih ograničenja koja određuju prihvatljivu veličinu modela, klasni model se može koristiti samostalno ili kao pomoćni model. U okviru ovog istraživanja razmatrani su različiti načini korišćenja klasnih modela jezika. Prvi korak predstavljala je linearna interpolacija osnovnog modela, odnosno modela reči, sa modelima baziranim na lemama i morfološkim klasama (Ostrogonac et al, 2012a). U ovom slučaju, rezultujući model se ne može nazvati

statističkim, s obzirom na činjenicu da se od lematskog i morfološkog modela dobijaju verovatnoće sekvenci lema, odnosno morfoloških klasa, a ne verovatnoće sekvenci reči. Ipak, linearnom kombinacijom izlaza pomenutih modela vrši se neka vrsta ublažavanja efekata strukture korpusa koji se koristi za obuku modela, što u praksi može biti od koristi, naročito kada se radi o primenama poput prepoznavanja govora, gde je uloga modela jezika vezana za poređenje verovatnoća hipoteza koje se dobijaju na izlazu akustičkog, odnosno leksičkog modela, a ne za određivanje verovatnoće neke konkretne sekvence. Kada su u pitanju primene poput korekcije grešaka u tekstovima, od koristi su upravo verovatnoće sekvenci morfoloških klasa, koje se mogu koristiti paralelno sa verovatnoćama sekvenci reči, što podrazumeva upravo paralelnu analizu izlaza osnovnog i klasnog modela, a ne njihovu kombinaciju (Ostrogonac, 2016). O istraživanjima vezanim za različite primene jezičkih modela biće više reči u poglavlju 5.

Kada su u pitanju morfološke klase reči, postoje određeni praktični problemi kod primene modela koji se baziraju na njima. Najočigledniji problem jeste činjenica da je broj klasa reči koje se mogu definisati na osnovu morfoloških informacija ograničen. Za neke primene, grupisanje reči može biti od koristi, ali ne ako je radikalna u toj meri da se desetine hiljada reči tretiraju kao jedna, što je slučaj kod morfoloških klasa. S druge strane, automatska klasterizacija Braunovim algoritmom omogućava kreiranje proizvoljnog broja klasa, ali je optimalan broj klasa teško odrediti, s obzirom na to da on zavisi i od veličine i od sadržaja tekstualnog korpusa, a ne samo od primene klasnog modela. S obzirom na činjenicu da Braunov algoritam (u originalnom obliku) funkcioniše tako što se iterativno vrši spajanje klasa, pri čemu se počinje od toga da je svaka reč klasa za sebe, kombinacija ovog algoritma sa grupisanjem na osnovu morfoloških obeležja nije izvodljiva. U okviru ovog istraživanja pokazano je da klasni modeli bazirani na morfološkim klasama daju bolje rezultate od klasnih modela baziranih na automatski izvedenim klasama, kada se primene u sistemu za prepoznavanje govora, pri čemu se broj automatski izvedenih klasa podešava tako da odgovara broju morfoloških klasa radi adekvatnog poređenja modela (Ostrogonac et al, 2018). Međutim, istovremeno se došlo do zaključka da je broj morfoloških klasa ipak premali za potrebe prepoznavanja govora, čak i za aplikacije koje bi se koristile na uređajima sa relativno oskudnim resursima (detalji ovog dela istraživanja će takođe biti prikazani u poglavlju 5). Stoga i dalje postoji potreba da se definiše način na koji bi se početni skup morfoloških klasa mogao proširiti na proizvoljan broj klasa.

Još jedan problem vezan za morfološke klasne modele jezika predstavlja činjenica da je pri korišćenju ovakvih modela u realnom vremenu potrebno vršiti automatsku morfološku anotaciju i klasifikaciju reči prema dobijenim morfološkim

informacijama. Ovaj proces je relativno vremenski zahtevan, a podrazumeva i korišćenje morfološkog rečnika, što za mnoge primene nije prihvatljivo. Jedno rešenje ovog problema jeste da se svaki pojavni oblik reči preslikava u tačno jednu morfološku klasu. Drugim rečima, ako se pri automatskoj morfološkoj anotaciji svaka reč posmatra izolovano, ona će biti anotirana uvek na isti način. Tako se može dobiti obostrano jednoznačna relacija između pojavnih oblika reči i morfoloških klasa. Ovo eliminiše potrebu za morfološkom anotacijom, kao i potrebu za korišćenjem rečnika. S druge strane, ovakvim grupisanjem gubi se značajna količina informacija. Primera radi, može se posmatrati sledeća rečenica:

Gore gore gore gore.

Automatskom morfološkom anotacijom pomoću ekspertskog sistema i klasifikacijom reči date rečenice, dobijaju se sledeće povratne informacije:

prid_kom_n_zr_m i_zr_nv_zaj_n_m prid_kom_g_zr_j i_zr_nv_zaj_g_j

i_zr_nv_zaj_n_m

U pitanju su, naravno identifikatori morfoloških klasa, na osnovu kojih se intuitivno, u izvesnoj meri, može uočiti kako je rečenica interpretirana. Naime, prva reč interpretirana je kao pridev u komparativu, pri čemu je padež nominativ, rod je ženski, a broj množina. Druga reč interpretirana je kao zajednička imenica ženskog roda u nominativu množine. Treća reč je interpretirana kao pridev u komparativu, pri čemu je padež nominativ, rod je ženski, a broj – jednina. Konačno, četvrta reč interpretirana je kao zajednička imenica ženskog roda, u genitivu jednine. Dakle, na osnovu konteksta, došlo se do interpretacije rečenice, koja je u ovom slučaju neispravna, ali činjenica je da je rečenica primer prilično retke situacije. Ipak, interpretacija koja je dobijena je značajno bliža ispravnoj od interpretacije koja se dobija bez analize konteksta. Naravno, i da je dobijena interpretacija ispravna, bila bi to samo jedna od više mogućih ispravnih interpretacija.

Kao što je ranije rečeno, efikasna alternativa morfološkoj anotaciji u realnom vremenu jeste kreiranje skupa relacija reč-klasa, gde se pri kreiranju tog skupa zapravo vrši morfološka anotacija svakog pojavnog oblika reči iz korpusa za obuku, ali tako što se uklanja kontekst. Na taj način, za rečenicu „*Gore gore gore gore.*“ povratne informacije bile bi:

i_zr_nv_zaj_n_m i_zr_nv_zaj_n_m i_zr_nv_zaj_n_m i_zr_nv_zaj_n_m

Dakle, u ovom slučaju bi rečenica bila interpretirana kao četiri uzastopne zajedničke imenice ženskog roda u nominativu množine. Ovde je, očigledno, došlo do veoma velikog gubitka informacija, kako bi se korišćenje klasnog modela učinilo efikasnijim. Međutim, u opštem slučaju situacija nije ni približno toliko dramatična,

a kao što će biti prikazano u poglavlju 5, čak i ovakvi klasni modeli pokazali su se uspešnijima od modela baziranih na klasama izvedenim pomoću Braunovog algoritma.

4.3 Hibridni modeli

Jedan od načina kombinovanja različitih jezičkih modela, preciznije modela reči, lema i morfoloških klasa, definisan je u okviru ovog istraživanja, a rezultujućem modelu dat je naziv hibridni model. Iako i dalje u idejnoj fazi, ovakav tip modela zaslužuje pažnju, jer pruža mogućnost minimalnog gubitka korisnih informacija koje su sadržane u tekstualnom korpusu, uz istovremeno tretiranje problema nedostajućih podataka za obuku modela. Struktura hibridnog modela za srpski jezik ima izvesnih sličnosti sa strukturom modela koji je kreiran za arapski jezik u okviru istraživanja opisanog u (Kirchhoff et al, 2006). U tom istraživanju, reči su predstavljane u vidu vektora obeležja, pri čemu su obeležja bila sam pojavni oblik, koren i vrsta reči. Model je predstavljao *N*-grame koje su sačinjavale sekvence skupova pomenutih obeležja i bio je veoma velikog obima. Za upotrebu ovakvog modela, definisan je *back-off* algoritam koji podrazumeva proveru postojanja *N*-grama koji sadrži informacije o svim obeležjima za svaku od reči koje pripadaju ulaznoj sekvenci, a ukoliko takav *N*-gram ne postoji, iterativno se odbacuje po jedno obeležje dok se ne pronađe *N*-gram koji odgovara ulaznim podacima. Pri tome nije definisana konkretna putanja za *back-off*, već je predloženo da se paralelno prolazi svim mogućim putanjama, a da se konačni rezultat dobije uprosečavanjem dobijenih rezultata.

Za srpski jezik, pored pojavnog oblika reči, dostupne su i informacije o osnovnim oblicima reči – lemama, kao i o morfološkim klasama. U slučaju npr. trigram modela, ukoliko se određeni trigram reči $w_1w_2w_3$ nije pojavio u korpusu za obuku, vrlo je moguće da se pojavio trigram kod koga su druga i treća reč identične originalnom *N*-gramu, dok je prva reč, recimo da je u pitanju imenica, imala isti osnovni oblik, samo je npr. broj bio različit, odnosno u jednom slučaju se radilo o jednini, a u drugom slučaju o množini. U opštem slučaju, umesto da se sa trigrama $w_1w_2w_3$ prelazi na bigrame w_2w_3, bolje je iskoristiti trigram $l_1w_2w_3$, gde će se umesto pojavnog oblika prve reči u sekvenci koristiti odgovarajući osnovni oblik. Za srpski jezik je definisana konkretna putanja za *back-off* proceduru, koja podrazumeva supstituciju sa kojom se počinje od najmanje važne reči (najudaljenijeg dela „istorije" – w_1), a zatim na prelazi na sledeću reč sve dok se ne dođe do reči koja se pojavljuje neposredno pre tekuće – za trigrame to bi zapravo bila već sledeća reč – w_2. Supstitucija na svakoj od reči koje čine istoriju vrši se u skladu sa poretkom kojim se podrazumeva da je pojavni oblik reči idealan slučaj, a ukoliko on ne postoji u određenom kontekstu, prelazi se najpre na lemu, pa na morfološku klasu, ukoliko ni lema nije pronađena. Ipak, ovakav model morao bi da

sadrži veliku količinu podataka te je potrebno razmotriti i načine smanjenja obima modela. O tehnikama potkresivanja osnovnih jezičkih modela i rezultatima za srpski jezik biće više reči u narednom poglavlju, ali kod hibridnih modela je situacija znatno složenija.

Kreiranje hibridnog modela za srpski jezik počinje od pripreme korpusa za obuku. U prethodnom poglavlju bilo je reči o korpusima koji sadrže reči, ili osnovne oblike reči, ili identifikatore morfoloških klasa. Za hibridni model, korpus se kreira tako što se svaka reč zamenjuje uređenom trojkom koju čine pojavni oblik, lema i identifikator morfološke klase. Primer sekvence iz korpusa za obuku ovakvog modela dat je u nastavku (običnim zagradama su obeležene leme, a uglastim zagradama identifikatori morfoloških klasa).

moramo (*morati*) [*g_mmf_nep_ne_prez_1_m*] *stići* (*stići*) [*g_mmf_pre_ne_ippsp*] *do* (*do*) [*pred_g*] *bolnice* (*bolnica*) [*i_zr_nv_zaj_g_j*]

Za razliku od istraživanja opisanog u (Kirchhoff et al, 2006), *N*-grami koji sačinjavaju rezultujući model neće se sastojati od sekvence pomenutih uređenih trojki, već će se za svaku sekvencu kreirati sve moguće kombinacije informacija koje čine te uređene trojke, kao što je prikazano na slici 13.

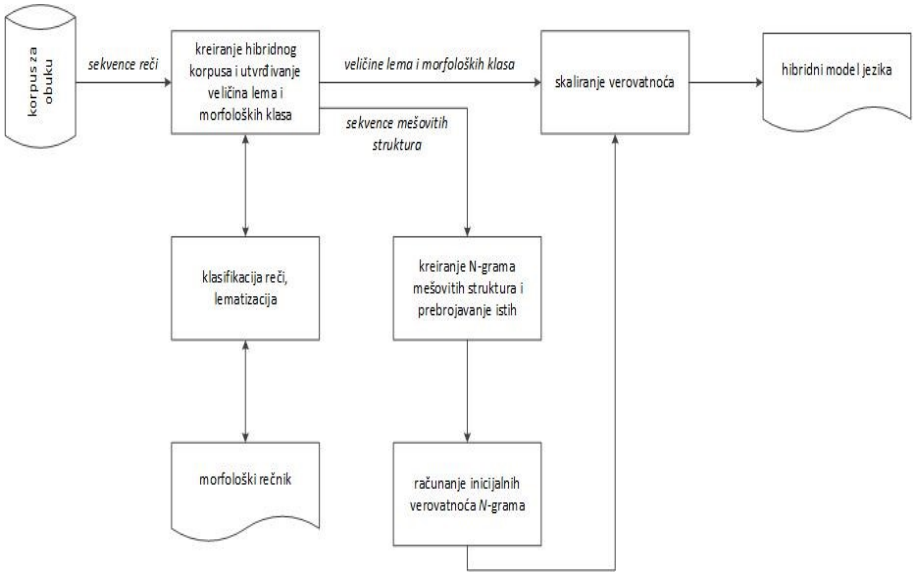
Kao što se vidi na slici, za svaki *N*-gram koji bi postojao u osnovnom jezičkom modelu, za hibridni model se generiše N^3 parametara. Naravno, ovakvi *N*-grami ne mogu se koristiti kao ravnopravni, jer verovatnoće reči, lema i morfoloških klasa predstavljaju odvojene raspodele. Ipak, uz skaliranje verovatnoća u zavisnosti od veličine lema, odnosno morfoloških klasa, u smislu broja reči koje se preslikavaju u svaku od njih, moglo bi se doći do određene ravnopravnosti. Detalji skaliranja verovatnoća u okviru ovog istraživanja nisu do kraja definisani, ali je jasno da se o ovome mora voditi računa. Na slici 14 prikazan je postupak obuke hibridnog modela za srpski jezik.

Ukoliko bi se definisao način adekvatnog skaliranja verovatnoća *N*-grama, tako da se *N*-grami koji sadrže različite strukture mogu ravnopravno koristiti, onda bi ovakav model mogao biti podvrgnut potkresivanju po potrebi, što je veoma značajno s obzirom na inicijalnu veličinu modela. Potkresivanjem bi se zapravo dobio model koji bi zadržao samo *N*-grame reči koji su se pojavljivali dovoljno često da se sa velikom sigurnošću može tvrditi da je verovatnoća njihovog pojavljivanja estimirana sa prihvatljivom tačnošću, a za ostale *N*-grame bi postojale alternative koje bi, na određenim pozicijama, umesto reči sadržale odgovarajuće leme ili klase. Ovakvi modeli bi najverovatnije eliminisali potrebu za *back-off* procedurom na *N*-grame nižeg reda, koji i sami po sebi unose mnogo problema u proces estimacije verovatnoće neke nove sekvence u fazi korišćenja modela.

Kada je reč o upotrebi hibridnog modela, jasno je da je za ulaznu sekvencu reči potrebno odrediti odgovarajuće sekvence lema i morfoloških klasa, nakon čega je potrebno kreirati listu *N*-grama mešovitih struktura kao što je prikazano na slici 13, na osnovu koje se onda vrši pretraga, počevši od *N*-grama koji sadrži samo reči, pa sve do *N*-grama koji sadrži samo identifikatore morfoloških klasa, ukoliko se ne pronađe nijedna druga kombinacija. Naravno, ako se ni poslednji *N*-gram sa liste ne bi pronašao, moralo bi se preći na niži red *N*-grama, ali kao što je ranije pomenuto, to bi se dešavalo veoma retko.

```
on je dokazao
on je (dokazati)
on je [g_nmmf_nep_np_per_mr_j]
on (biti) dokazao
on (biti) (dokazati)
on (biti) [g_nmmf_nep_np_per_mr_j]
on [g_je] dokazao
on [g_je] (dokazati)
on [g_je] [g_nmmf_nep_np_per_mr_j]
(on) je dokazao
(on) je (dokazati)
(on) je [g_nmmf_nep_np_per_mr_j]
(on) (biti) dokazao
(on) (biti) (dokazati)
(on) (biti) [g_nmmf_nep_np_per_mr_j]
(on) [g_je] dokazao
(on) [g_je] (dokazati)
(on) [g_je] [g_nmmf_nep_np_per_mr_j]
[i_mr_nv_n_j] je dokazao
[i_mr_nv_n_j] je (dokazati)
[i_mr_nv_n_j] je [g_nmmf_nep_np_per_mr_j]
[i_mr_nv_n_j] (biti) dokazao
[i_mr_nv_n_j] (biti) (dokazati)
[i_mr_nv_n_j] (biti) [g_nmmf_nep_np_per_mr_j]
[i_mr_nv_n_j] [g_je] dokazao
[i_mr_nv_n_j] [g_je] (dokazati)
[i_mr_nv_n_j] [g_je] [g_nmmf_nep_np_per_mr_j]
```

Slika 13. Kreiranje *N*-grama za hibridni model na osnovu sekvence reči dužine 3



Slika 14. Proces obuke hibridnog modela za srpski jezik

POGLAVLJE 5

PRIMENE JEZIČKIH MODELA

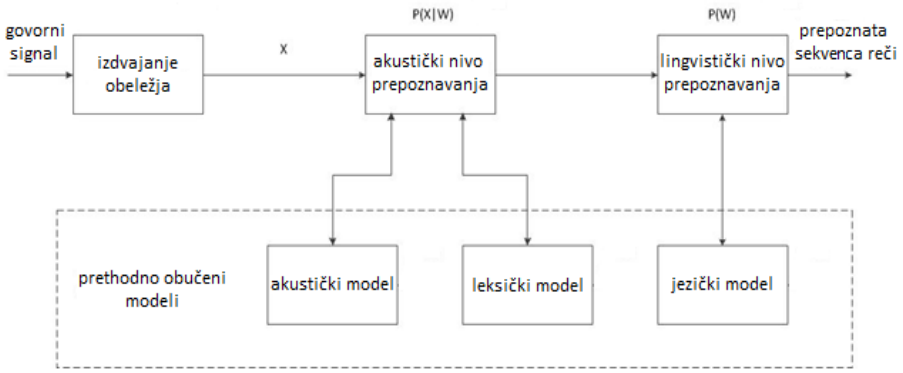
U prethodnim poglavljima opisano je stanje u oblasti jezičkog modelovanja, prikazan je proces prikupljanja i pripreme resursa za obuku jezičkih modela za srpski jezik i predstavljen je način tretiranja problema nedovoljne količine podataka za obuku modela kroz definisanje različitih struktura jezičkih modela. Čitav ovaj proces pratila su i istraživanja vezana za primenu jezičkih modela u govornim i jezičkim tehnologijama, koja su za cilj imala utvrđivanje najpovoljnijih tehnika obuke, adaptacije veličine, pa čak i evaluacije modela srpskog jezika, za pojedine primene ili uopšte. U ovom poglavlju, biće predstavljeni glavni rezultati pomenutih istraživanja, a biće predstavljeni i pravci daljeg razvoja za svaku od tehnologija koje su bile predmet ove teze.

5.1 Automatsko prepoznavanje govora

Tehnologija automatskog prepoznavanja govora razvijana je tokom protekle dve decenije na Fakultetu tehničkih nauka u Novom sadu, u saradnji sa preduzećem *AlfaNum* (Janev et al, 2010), uporedo sa razvojem sinteze govora na osnovu teksta (Sečujski et al, 2007). Povećanje tačnosti automatskog prepoznavanja govora predstavljalo je inicijalnu motivaciju za razvoj jezičkih modela za srpski jezik. Stoga je najviše eksperimenata urađeno upravo sa ciljem da se utvrde standardi za kreiranje jezičkih modela za potrebe prepoznavanja govora na malim i velikim rečnicima.

Na slici 15 šematski je prikazana standardna struktura sistema za automatsko prepoznavanje govora. Iz govornog signala se, na početku procesa, izdvajaju određena akustička obeležja. Ova obeležja su vezana za oblik spektralne obvojnice signala, osnovnu učestanost, energiju i uglavnom se, pored statičkih obeležja, uzimaju u obzir i dinamička, odnosno analizira se promena pomenutih obeležja u vremenu. Na osnovu prethodno obučenih akustičkih modela, dobijaju se hipoteze o sekvencama fonema. Ove hipoteze podvrgavaju se evaluaciji koja se vrši pomoću leksičkog modela. Izlaz leksičkog modela, odnosno ulazni podaci za jezički model, jeste lista najverovatnijih hipoteza od kojih svaka predstavlja sekvencu reči. U okviru ovog sistema, dakle, uloga modela jezika jeste evaluacija ulaznih hipoteza, nakon koje se dolazi do najverovatnije sekvence reči, koja predstavlja izlaz ASR sistema.

Značaj jezičkog modela u prepoznavanju govora u odnosu na akustičke modele (i leksičke) moguće je podešavati težinskim koeficijentima. Težinski koeficijenti najčešće se određuju eksperimentisanjem sa različitim vrednostima uz analizu rezultujućih grešaka prepoznavanja. Međutim, u ovom procesu značajno pomaže prethodna evaluacija samih modela, u meri u kojoj je to moguće izvršiti.



Slika 15. Sistem za automatsko prepoznavanje govora

Evaluacija jezičkih modela, kada se oni posmatraju kao izolovani elementi, najčešće se vrši računanjem perpleksnosti, o kojoj je bilo reči u uvodnom poglavlju. Iako se pokazalo da perpleksnost ne mora nužno da odgovara doprinosu koji jezički model može da ima u vidu smanjenja greške pri prepoznavanju govora, jer taj doprinos zavisi od mnoštva faktora (domen primene, kvalitet akustičkih modela, reprezentativnost teksta koji se koristi za evaluaciju modela računanjem perpleksnosti itd.), ova vrsta evaluacije ipak predstavlja dobar početni korak u utvrđivanju kvaliteta jezičkog modela.

U nastavku ovog odeljka biće izloženi rezultati eksperimenata čiji je cilj bio utvrđivanje pravila za kreiranje jezičkih modela za potrebe prepoznavanja govora za različite aplikacije za srpski jezik, koje nameću specifične uslove.

Utvrđivanje uticaja veličine korpusa za obuku na kvalitet modela jezika

Nakon što su pripremljeni tekstualni korpusi za srpski jezik, urađen je eksperiment sa ciljem da se utvrdi uticaj veličine korpusa na kvalitet jezičkih modela i, što je naročito važno, kvalitet modela koji se dobija kada se upotrebi kompletan korpus za obuku (Ostrogonac et al, 2012c). U slučaju upotrebe u prepoznavanju govora iz tekstualnih sadržaja korišćenih u eksperimentima

izostavljani su znaci interpunkcije. Posmatrani su modeli bazirani na originalnom tekstu, kao i modeli bazirani na lemama i morfološkim klasama. Za svaki od pomenuta tri tipa modela obučavano je po 17 instanci, tako što je inkrementalno povećavan procenat korpusa za obuku koji je korišćen. Navedenih 17 modela bilo je podeljno u dve grupe. Prva grupa od 8 modela obučavana je na veoma maloj količini podataka (okvirno između 1 i 10% od ukupnog korpusa), uz konsekventno mali inkrement, kako bi se mogle pribaviti detaljnije informacije o kvalitetu pojedinih tipova modela u situacijama kada je dostupno vrlo malo podataka za obuku. Druga grupa sadržala je 9 modela i tu je inkrement bio značajno veći (oko 10% od ukupnog korpusa). Kada se govori o ukupnom korpusu, kod ovog eksperimenta se misli na korpus za novinarski stil. Ovaj korpus je izabran zbog toga što je najvećeg obima, ali i zbog toga što je novinarski stil, u nedostatku korpusa za razgovorni stil, najpogodniji za kreiranje modela opšte namene, koji su od interesa u eksperimentima poput ovog, gde je cilj određivanje opštih pravila za modelovanje određenog jezika. Obuka modela vršena je pomoću alata *SRILM*, a kreirani su *Katz* trigram modeli. U tabeli 3 prikazani su podaci o veličinama korpusa koji su korišćeni za obuku modela, pri čemu je prva grupa modela označena sa M01-M08, dok je druga grupa modela označena sa M1-M9. Ovde je važno napomenuti da broj definisanih morfoloških nije inicijalno bio 1124, jer u početku nisu korišćene određene informacije vezane za imenice pri grupisanju, te je broj klasa koje su se pojavile u korpusu u okviru ovog eksperimenta stoga manji od broja koji je prikazan u poglavlju 3. Takođe, i sam korpus je u međuvremenu proširivan, iako ne značajno.

Svi modeli evaluirani su računanjem perpleksnosti na posebnom delu korpusa koji je odvojen da služi kao test skup. Ovaj deo predstavljao je oko 10% ukupnog korpusa. Rezultati evaluacije prikazani su u tabeli 4. Pri tome, perpleksnosti za modele bazirane na lemama i morfološkim klasama računate su na odgovarajućim nivoima, a ne na način na koji se računa perpleksnost klasnih modela koji se koriste za računanje verovatnoća sekvenci reči. Proračun je izvršen na ovaj način zbog toga što je za potrebe ispitivanja uticaja veličine korpusa za obuku na kvalitet modela bilo od interesa saznati koliki je korpus dovoljan da se ovakvi modeli obuče, odnosno u kom koraku pri povećanju korpusa će doći do stagnacije perpleksnosti na nivou reči, lema, odnosno morfoloških klasa, u zavisnosti od toga o kakvim modelima se radi. Pored eksperimenta za koji su rezultati prikazani u tabeli 4, urađena je i evaluacija modela na test skupu koji je prethodno modifikovan tako što je redosled reči u svakoj od rečenica izmenjen na slučajan način. Rezultati ovog eksperimenta prikazani su u tabeli 5, a od značaja su za računanje mere kvaliteta modela koja je definisana u okviru ovog istraživanja kao količnik perpleksnosti koja se dobija na tekstu sa slučajnim rasporedom reči i perpleksnosti koja se dobija na autentičnom tekstu. Ova mera nazvana je koeficijent diskriminacije (*KD*), i predstavlja sposobnost modela da razlikuje

smislen tekst od sekvence na slučaj odabranih reči. Pomoću ove mere mogu se direktno porediti modeli reči sa modelima lema i morfoloških klasa. Ako se uzmu u obzir izraz za računanje perpleksnosti modela jezika (1.5) i izraz za računanje verovatnoće sekvence reči na osnovu klasnog modela jezika (4.1), koeficijent diskriminacije računa se na sledeći način:

$$KD = \frac{ppl_s}{ppl_a} = \frac{\sqrt[m]{\frac{1}{\prod_{i=1}^{m} P_S(w_i|w_{i-N+1}\dots w_{i-1})}}}{\sqrt[m]{\frac{1}{\prod_{i=1}^{m} P_a(w_i|w_{i-N+1}\dots w_{i-1})}}} = \sqrt[m]{\frac{\prod_{i=1}^{m} P_a(w_i|w_{i-N+1}\dots w_{i-1})}{\prod_{i=1}^{m} P_S(w_i|w_{i-N+1}\dots w_{i-1})}} =$$

$$\sqrt[m]{\frac{\prod_{i=1}^{m} P(w_i|c_i)P_a(c_i|c_{i-N+1}\dots c_{i-1})}{\prod_{i=1}^{m} P(w_i|c_i)P_S(c_i|c_{i-N+1}\dots c_{i-1})}}. \qquad (5.1)$$

U izrazu (5.1), ppl_s predstavlja vrednost perpleksnosti na tekstu sa slučajnim redosledom reči na nivou rečenice, dok, ppl_a predstavlja vrednost perpleksnosti na autentičnom tekstu. Analogno tome, P_s predstavljaju verovatnoće sekvenci koje se pojavljuju u tekstu sa slučajnim redosledom reči, dok P_a predstavljaju verovatnoće sekvenci koje se pojavljuju u autentičnom tekstu. Iako se u autentičnom tekstu ne pojavljuju isti *N*-grami koji se pojavljuju u tekstu sa slučajnim redosledom reči, proizvodi verovatnoća da pojedine reči pripadaju određenim klasama su jednaki u brojiocu i imeniocu. Iz toga sledi da koeficijent diskriminacije ne zavisi od strukture modela jezika, odnosno od načina grupisanja reči.

Vrednosti *KD* prikazane su u tabeli 6, a na slikama 16 i 17 ove vrednosti prikazane su i grafički, posebno za prvu i drugu grupu modela. Zbog načina na koji se računaju vrednosti koeficijenta diskriminacije, moguće je pomoću ove mere direktno porediti modele različitih struktura. O ovome će biti više detalja u narednom delu ovog odeljka.

modeli	ukupno reči ($x10^6$)	reči ($x10^3$)	lema ($x10^3$)	morfoloških klasa
M01	0,09	20,21	11,37	506
M02	0,20	32,36	17,00	552
M03	0,40	47,09	23,58	586
M04	0,58	57,29	27,97	605
M05	0,91	73,70	35,24	628
M06	1,14	84,62	40,70	642
M07	1,37	93,86	44,18	646
M08	1,60	101,71	47,56	651
M1	1,75	106,78	49,94	652
M2	3,52	149,82	69,79	681
M3	5,28	181,08	84,60	700
M4	6,95	205,78	96,61	710
M5	8,59	226,66	107,05	716
M6	10,26	245,80	116,65	720
M7	11,91	262,90	125,33	727
M8	13,56	278,97	133,47	731
M9	15,24	293,65	140,86	733

Tabela 3. Podaci o korpusima za obuku modela koji su korišćeni za utvrđivanje uticaja veličine korpusa na kvalitet modela

Ono što je zanimljivo uočiti na slici 16 jeste činjenica da klasni model, u situacijama kada je dostupna vrlo mala količina podataka, daje značajno bolje rezultate od modela baziranih na lemama i rečima. Pri tome, modeli reči u odnosu na modele lema daju bolje rezultate, a ta razlika je proporcionalna količini podataka za obuku i, za male korpuse, kakvi su korišćeni za obuku prve grupe modela, nije značajna. Na slici 17 vidi se da je oko 20% postojećeg korpusa za novinarski stil dovoljno sa se modeli bazirani na morfološkim klasama adekvatno obuče. Drugim rečima, daljim povećanjem korpusa za obuku ne dobija se značajno na kvalitetu modela koji su bazirani na morfološkim klasama. S obzirom na relativno mali broj klasa koje su definisane na osnovu morfoloških informacija,

ovakvi modeli imaju, zapravo, vrlo ograničen potencijal kada je u pitanju diskriminativna moć.

modeli	reči	leme	m. klase
M01	599,20	434,57	31,26
M02	634,27	409,65	28,82
M03	642,43	392,58	27,25
M04	639,03	378,45	26,58
M05	615,31	352,23	25,74
M06	599,16	338,65	25,35
M07	548,29	310,22	25,01
M08	532,46	298,91	24,80
M1	484,72	275,47	24,59
M2	389,86	227,64	23,82
M3	337,00	202,23	23,42
M4	305,91	186,52	23,21
M5	288,76	177,94	23,08
M6	276,02	171,51	22,97
M7	267,99	167,42	22,91
M8	259,70	163,78	22,85
M9	239,70	153,29	22,75

Tabela 4. Vrednosti perpleksnosti za modele reči, lema i morfoloških klasa na autentičnom tekstu

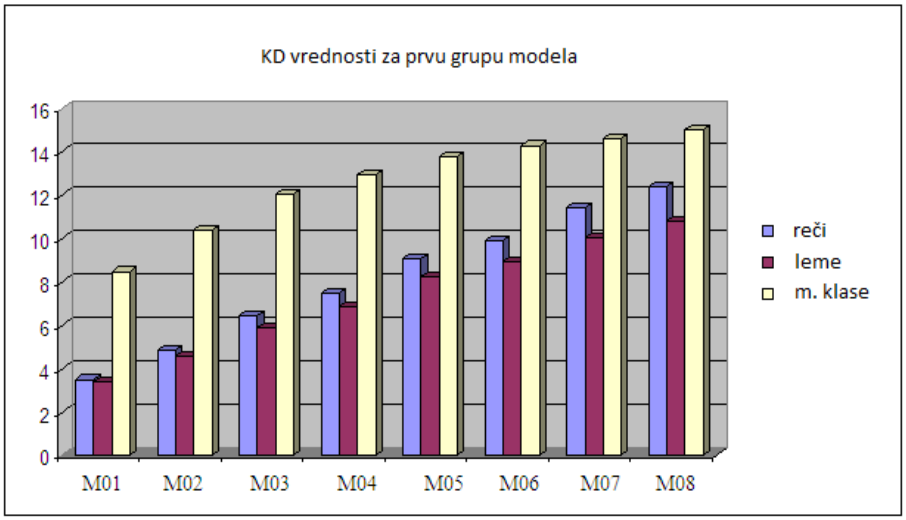
Dakle, za veće količine podataka za obuku bolje rezultate postizaće modeli reči i lema. Na slici 17 se, ipak, vidi da povećanjem količine podataka za obuku ne dolazi do zasićenja u smislu porasta diskriminativne moći modela. To potvrđuje pretpostavku da ni kompletan korpus nije dovoljan da se maksimalno iskoristi potencijal modela jezika. Međutim, iako se pri velikim količinama podataka modeli reči pokazuju kao najbolji, to ne znači da modeli lema i morfoloških klasa ne sadrže specifične informacije kojima bi se moglo doprineti kvalitetu osnovnog modela (modela reči).

modeli	reči	leme	m. klase
M01	2.097,56	1.475,91	265,30
M02	3.076,31	1.877,92	299,68
M03	4.154,21	2.310,44	328,54
M04	4.791,33	2.579,93	343,65
M05	5.583,10	2.898,68	354,75
M06	5.939,57	3.012,17	362,48
M07	6.265,49	3.123,36	365,04
M08	6.598,95	3.223,02	372,54
M1	6.815,13	3.290,24	375,23
M2	8.238,23	3.844,57	394,63
M3	9.149,97	4.150,64	407,19
M4	9.850,25	4.393,96	413,28
M5	10.412,40	4.598,64	417,74
M6	10.884,60	4.761,98	421,97
M7	11.317,80	4.913,70	424,37
M8	11.670,80	5.053,69	427,63
M9	12.023,30	5.182,77	430,03

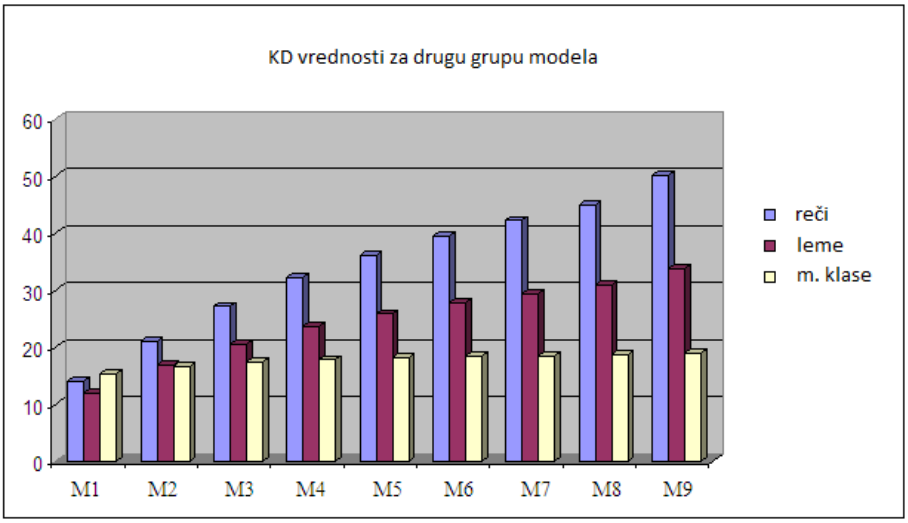
Tabela 5. Vrednosti perpleksnosti za modele reči, lema i morfoloških klasa na tekstu sa slučajnim rasporedom reči na nivou rečenice

modeli	reči	leme	m. klase
M01	3,50	3,40	8,49
M02	4,85	4,58	10,40
M03	6,47	5,89	12,06
M04	7,50	6,86	12,93
M05	9,07	8,23	13,78
M06	9,91	8,92	14,30
M07	11,42	10,07	14,60
M08	12,39	10,78	15,02
M1	14,06	11,94	15,26
M2	21,13	16,89	16,57
M3	27,15	20,52	17,38
M4	32,20	23,65	17,80
M5	36,06	25,84	18,10
M6	39,43	27,76	18,37
M7	42,23	29,35	18,52
M8	44,94	30,86	18,71
M9	50,16	33,81	18,90

Tabela 6. Vrednosti koeficijenta diskriminacije za modele reči, lema i morfoloških klasa



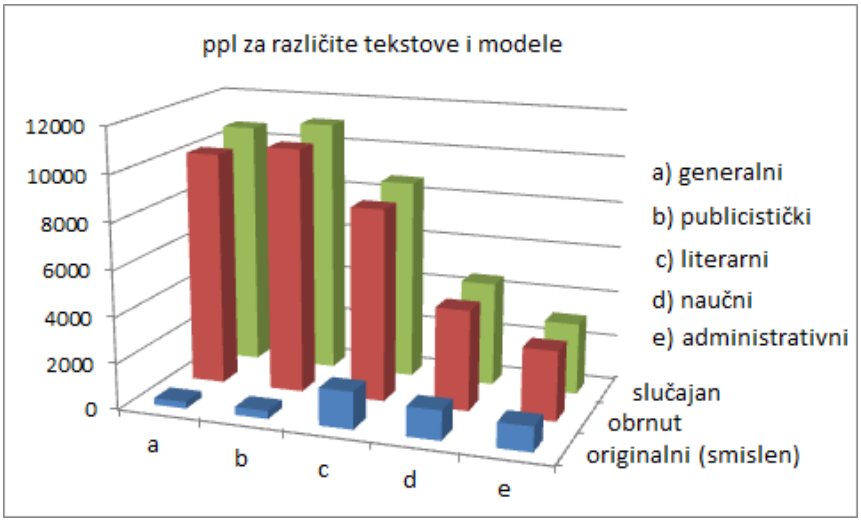
Slika 16. Vrednosti koeficijenta diskriminacije za prvu grupu modela reči, lema i morfoloških klasa



Slika 17. Vrednosti koeficijenta diskriminacije za drugu grupu modela reči, lema i morfoloških klasa

Diskriminativne mogućnosti različitih tipova modela

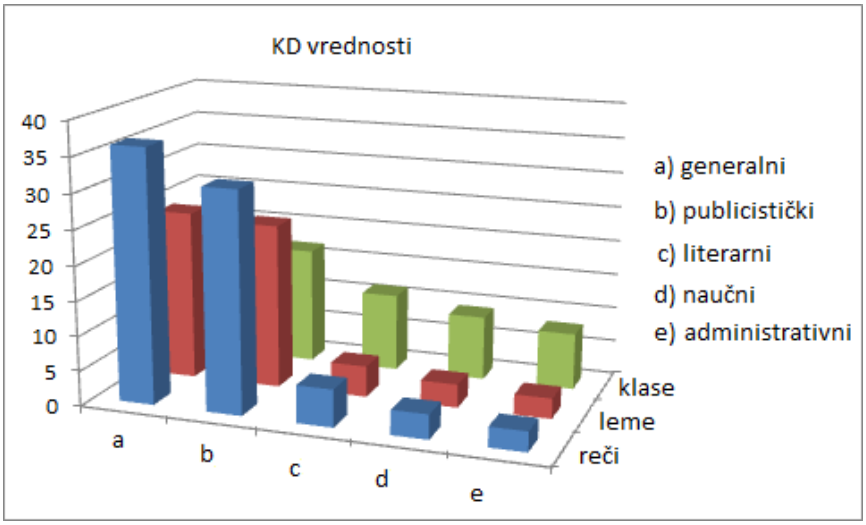
Zaključci do kojih se došlo istraživanjem vezanim za zavisnost kvaliteta modela jezika od veličine korpusa nametnuli su pitanje koliki je kvalitet modela koji se može postići za pojedine funkcionalne stilove, kao i pitanje koliki je kvalitet modela koji se dobija obukom nad združenim korpusima. Na slikama 18 i 19 prikazani su rezultati ovog ispitivanja, pri čemu su modeli koji su obučavani na kompletnom tekstualnom korpusu, odnosno na združenim korpusima za pojedine funkcionalne stilove, nazvani generalnim modelima. Svi modeli su i u ovom ispitivanju *Katz* trigram modeli (Ostrogonac et al, 2012b).



Slika 18. Vrednosti perpleksnosti za osnovne modele (modele reči) koji su obučavani na korpusima za pojedine funkcionalne stilove, kao i za model obučen na združenim korpusima, pri čemu je evaluacija vršena na autentičnom tekstu, tesktu sa obrnutim redosledom reči na nivou rečenice i tekstu sa slučajnim rasporedom reči na nivou rečenice

Slika 18 prikazuje rezultate koji su dobijeni za modele obučavane na osnovnim korpusima, odnosno korpusima koji sadrže reči. Evaluacija je vršena na autentičnom tekstu, kao i na tekstu sa obrnutim i na tekstu sa slučajnim rasporedom reči na nivou rečenice. Ovde je zanimljivo primetiti da je za obrnut redosled reči, u odnosu na slučajan redosled reči na nivou rečenice, za sve funkcionalne stilove dobijena manja perpleksnost na podacima za testiranje, iako razlika nije drastična. Može se pretpostaviti da bi se slučajnim raspoređivanjem reči na nivou kompletnog teksta koji se koristi za evaluaciju dobila još veća vrednost perpleksnosti. Razlika te vrednosti i vrednosti koja je dobijena u slučaju

kada su reči raspoređene na slučajan način na nivou rečenice predstavlja svojevrsnu meru količine informacija koju nosi kontekst izvan granica određene rečenice. Mogućnost utvrđivanja ovakve mere za različite tipove tekstualnih sadržaja bila bi od velike koristi u mnogim primenama govornih i jezičkih tehnologija, te stoga predstavlja jednu od tema budućih istraživanja. Na slici 19 uporedo su prikazane vrednosti koeficijenata diskriminacije za modele obučene na korpusima za pojedinačne funkcionalne stilove i za modele obučene na celokupnom tekstualnom korpusu, pri čemu su evaluirani modeli reči, lema i morfoloških klasa. Ovaj eksperiment je pokazao da samo za novinarski (publicistički) stil postoji dovoljno podataka da model reči, kada se koristi kao samostalan, postigne bolje rezultate od morfološkog klasnog modela. Za ostale funkcionalne stilove to nije slučaj, kako zbog same veličine korpusa, tako i zbog prirode pojedinih stilova. Ono što je takođe važno primetiti jeste činjenica da se združivanjem svih korpusa koji odgovaraju pojedinim funkcionalnim stilovima postiglo značajno poboljšanje u odnosu na model koji je obučavan samo na korpusu za novinarski stil, iako dodati sadržaji sadrže drugačije rečenične konstrukcije, a postoje i značajne razlike u rečnicima. Još jedan zanimljiv detalj koji je uočljiv na slici 19 jeste ponašanje modela baziranog na lemama kod literarnog, naučnog i administrativnog stila. Naime, za male korpuse za obuku, ovakvi modeli daju lošije rezultate od modela reči i od modela morfoloških klasa, što govori o tome da je grupisanje reči po morfološkim obeležjima značajno bolji pristup od grupisanja po značenju u situacijama kada je potrebno kreirati modele na osnovu male količine podataka.



Slika 19. Vrednosti koeficijenta diskriminacije za drugu grupu modela reči, lema i morfoloških klasa

Poređenje morfoloških i automatski izvedenih klasa

U prethodnim poglavljima pominjan je Braunov algoritam, kojim je moguće automatski klasterizovati reči na osnovu statistika tekstualnog korpusa. Eksperimenti koji su rađeni za engleski jezik pokazali su da se pomoću ovog algoritma mogu kreirati klase reči koje su slične po značenju (Brown et al, 1992). Međutim, optimalan broj klasa ne određuje se automatski, pa je potrebno analizirati rezultate klasterizacije, kako bi se utvrdilo kada je potrebno zaustaviti proces. S druge strane, morfološke klase koje su definisane za srpski jezik ne obezbeđuju grupisanje reči po značenju. Pored toga, broj morfoloških klasa je fiksan. U istraživanju koje je opisano u (Ostrogonac et al, 2018) pretpostavka je, ipak, bila da se grupisanjem reči na osnovu morfoloških obeležja dobijaju adekvatniji modeli jezika od modela koji se dobijaju automatskim izvođenjem klasa. U prilogu 4 dati su primeri sadržaja nekih od morfoloških i automatski izvedenih klasa. Ovo istraživanje bilo je od naročitog značaja za primenu jezičkih modela u automatskom prepoznavanju govora, te su modeli, pored evaluacije računanjem perpleksnosti na test skupu podataka, evaluirani i po doprinosu smanjenju greške prepoznavanja govora na nivou reči. Može se očekivati da kod prepoznavanja govora samo značenje reči ne igra važnu ulogu, s obzirom na to da se hipoteze koje se dobijaju na izlazu leksičkog modela najčešće razlikuju po morfološkim obeležjima pojedinih reči, naravno ako su akustički modeli relativno dobro obučeni.

Evaluacija modela računanjem perpleksnosti rađena je na statističkim *N*-gram modelima, koji su obučavani pomoću alata *SRILM*, kao i na modelima baziranim na rekurentnim neuronskim mrežama, koji su obučavani pomoću alata *RNNLM*. Za obuku je korišćeno 90% postojećih korpusa za novinarski, literarni, naučni i administrativni stil. Morfološki modeli su obučavani na korpusima koji su dobijeni morfološkom anotacijom bez analize konteksta, kako bi se dobilo obostrano jednoznačno preslikavanje reči u morfološke klase. Od značaja je bilo ispitati kvalitet ovakvih morfoloških modela s obzirom na to da su oni pogodniji za praktičnu upotrebu. Rezultati za statističke *N*-gram modele prikazani su u tabeli 7. Za svaki funkcionalni stil naznačena je veličina rečnika *C*, odnosno broj morfoloških klasa koje su se pojavile u korpusu za obuku (pri morfološkoj anotaciji bez analize konteksta). Slovom *M* označeni su morfološki modeli, a slovom *U* modeli bazirani na automatski izvedenim klasama. Vrednosti perpleksnosti u tabeli 7 odnose se na klasni nivo. Na nivou reči, ove vrednosti se dobijaju korišćenjem verovatnoća pripadanja pojedinih reči odgovarajućim klasama. Ove vrednosti perpleksnosti su date u tabeli 8.

funkcionalni stil	tip modela	bigram	trigram	kvadrigram
administrativni (C = 443)	U	55,49	41,50	38,76
	M	31,05	22,55	20,68
literarni (C = 828)	U	125,94	108,60	107,38
	M	64,52	52,14	50,93
naučni (C = 646)	U	124,32	110,61	109,23
	M	43,74	36,23	35,68
novinarski (C = 836)	U	77,72	54,29	47,40
	M	43,80	32,31	28,35

Tabela 7. Rezultati evaluacije statističkih N-gram modela računanjem perpleksnosti na klasnom nivou (C - veličina rečnika, U – modeli bazirani na automatski izvedenim klasama, M – morfološki modeli)

funkcionalni stil	tip modela	bigram	trigram	kvadrigram
administrativni (C = 443)	U	1.052,64	816,64	762,74
	M	1.250,72	912,68	834,86
literarni (C = 828)	U	8.089,15	6.974,93	6.896,57
	M	3.629,93	2.949,15	2.877,67
naučni (C = 646)	U	6.596,25	5.868,64	5.795,55
	M	3.268,83	2.727,51	2.679,21
novinarski (C = 836)	U	9.235,24	6.450,65	5.631,81
	M	7.744,16	5.753,14	5.057,89

Tabela 8. Rezultati evaluacije statističkih N-gram modela računanjem perpleksnosti na nivou reči (C - veličina rečnika, U – modeli bazirani na automatski izvedenim klasama, M – morfološki modeli)

Podaci iz tabele 7 pokazuju da su morfološki modeli, bez obzira na funkcionalni stil, značajno bolji od modela baziranih na automatski izvedenim klasama, ali na klasnom nivou. Pitanje je, međutim, kakve su raspodele reči po klasama i kakav je kvalitet modela kada se oni koriste za predviđanje sledeće reči na osnovu konteksta. Na ovo pitanje daju podaci iz tabele 8. Vidi se da i na nivou reči morfološki modeli postižu značajno bolje rezultate, osim u slučaju administrativnog stila. Korpus za administrativni stil je značajno manji od ostalih korpusa, a sam stil ne odlikuje morfološka šarenolikost u meri u kojoj je ona izražena kod ostalih stilova, što može objasniti ovakav rezultat. Međutim, vrednosti perpleksnosti svih modela su veoma velike. Ovo navodi na zaključak da je broj morfoloških klasa isuviše mali da bi ovakvi modeli mogli samostalno adekvatno modelovati jezik. Ovo nije iznenađujući zaključak, s obzirom na to da su morfološki modeli i predviđeni za korišćenje u situacijama kada su ograničenja po pitanju veličine modela ekstremna, ili u situacijama kada resursi nisu problem, u kojima je pogodno koristiti ih kao pomoćne modele. S druge strane, grupisanje reči na osnovu morfoloških informacija može se posmatrati kao inicijalno grupisanje, na osnovu kojeg bi se mogao definisati veći broj klasa, u zavisnosti od veličine korpusa za obuku i drugih parametara. Na taj način bi se moglo doći do optimalnih modela za pojedine primene.

Kada su u pitanju modeli jezika bazirani na rekurentnim neuronskim mrežama, za njihovu obuku korišćene su preporuke koje su date u (Mikolov et al, 2011b) za korpuse „srednje veličine“, kako su u okviru tog ispitivanja klasifikovani korpusi sličnog obima kao što su korpusi za srpski. Ove preporuke podrazumevaju veličinu skrivenog sloja mreže 500, veličinu klasnog izlaznog sloja 400 i obuku pomoću *BPTT* algoritma u 10 epoha u blok režimu. Radi ilustracije, primer poziva alata *RNNLM* radi obuke modela sa pomenutim parametrima dat je u nastavku:

```
rnnlm.exe -train Obuka.txt -valid Test.txt -rnnlm
Model.txt -hidden 500 -rand-seed 1 -debug 0 -bptt-block 10 -class
400 1>log.txt
```

S obzirom na to da se radi o preporukama, a ne o optimizovanim parametrima za konkretan zadatak, kao i zbog činjenice da su dobijeni modeli bazirani na neuronskim mrežama mnogo obimniji od statističkih *N*-gram modela, ove dve vrste modela nema smisla direktno porediti, ali cilj ovog eksperimenta bio je samo da se ispita ponašanje morfološkog i automatskog grupisanja reči, odnosno da se ona uporede i kod modela koji su bazirani na rekurentnim neuronskim mrežama. Rezultati evaluacije računanjem perpleksnosti za RNN modele prikazani su u tabeli 9.

funkcionalni stil	ppl za M-modele	ppl za U-modele
administrativni (C = 443)	1.389,87	1.636,44
literarni (C = 828)	4.065,68	10.500,93
naučni (C = 646)	3.994,52	10.412,45
novinarski (C = 836)	6.273,07	11.543,21

Tabela 9. Rezultati evaluacije RNN modela računanjem perpleksnosti na nivou reči (C - veličina rečnika, U – modeli bazirani na automatski izvedenim klasama, M – morfološki modeli)

Test skup za evaluaciju RNN modela podeljen je na validacioni i evaluacioni deo, od kojih svaki sadrži po 5% ukupnog korpusa za svaki od funkcionalnih stilova. Korpusi za obuku, kao i veličine rečnika, odnosno brojevi klasa po stilovima, identični su onima koji su korišćeni za obuku N-gram modela. Ipak, čini se da RNN izraženije favorizuje grupisanje reči na osnovu morfoloških informacija. Zapravo, na osnovu vrednosti perpleksnosti za modele bazirane na automatski izvedenim klasama, čini se da ovakav pristup grupisanju reči nije adekvatan kada se u jezičkom modelovanju koristi paradigma neuronskih mreža.

Evaluacija modela računanjem greške prepoznavanja govora na nivou reči izvršena je pomoću ASR sistema koji je razvijen u preduzeću *AlfaNum* (Pakoci et al, 2017), a baziran na alatu za prepoznavanje govora *Kaldi* (Povey et al, 2011). Za testove je korišćena govorna baza koja sadrži 18 h snimaka, a materijale izgovara 26 različitih govornika. Baza sadrži 13.000 rečenica, oko 160.000 reči, a veličina rečnika je oko 27.000. Snimci su studijskog kvaliteta, a većina snimaka predstavlja audio-knjige, koje karakteriše literarni funkcionalni stil. Snimci su jednokanalni PCM zapisi, frekvencija odabiranja signala je 16 kHz, a broj bita kojim se koduju pojedini odbirci je 16. Za akustičke modele korišćena je duboka neuronska mreža sa vremenskim kašnjenjem (eng. *Time Delay Deep Neural Network* – *TDNN*). Obuka je rađena na bazi čiji je deo opisan u prethodnom pasusu. Ova baza ukupno sadrži oko 200 h snimaka, uglavnom sačinjenih od audio-knjiga (Suzić et al, 2014). Zbog nedostatka softverske podrške za evaluaciju klasnih modela baziranih na neuronskim mrežama, evaluacija po WER obuhvatila je samo statističke N-gram modele. Značaj jezičkog modela, odnosno težinski faktor pri ocenjivanju hipoteza,

variran je u nekoliko preliminarnih eksperimenata kako bi se utvrdila optimalna vrednost ovog parametra. Rezultati su prikazani u tabeli 10.

funkcionalni stil	tip modela	bigram WER	trigram WER	kvadrigram WER
administrativni (C = 443)	U	58,14	58,31	58,32
	M	57,45	57,61	57,65
literarni (C = 828)	U	30,56	30,66	31,32
	M	31,15	31,30	32,07
naučni (C = 646)	U	45,64	45,58	45,55
	M	42,81	43,14	43,44
novinarski (C = 836)	U	40,59	41,09	41,29
	M	35,66	36,29	36,82

Tabela 10. Rezultati evaluacije statističkih N-gram modela računanjem WER (C - veličina rečnika, U – modeli bazirani na automatski izvedenim klasama, M – morfološki modeli)

Važno je napomenuti da je procenat reči koje su se pojavile u test skupu, a koje nisu viđene u korpusu za obuku (eng. *Out-of-Vocabulary – OOV*), za neke od stilova bio izuzetno visok. Za administrativni stil procenat OOV reči bio je čak 36,37%, za literarni 3,93%, za naučni 19,3%, a za novinarski 4,62%. Ovo objašnjava generalno veoma visoke vrednosti WER. Za naučni i novinarski stil, morfološki modeli su se pokazali kao značajno uspešniji, za administrativni stil morfološki modeli su u blagoj prednosti, dok su se za literarni stil modeli bazirani na automatski izvedenim klasama pokazali kao marginalno adekvatniji. Uočljiv detalj jeste malo povećanje WER do kojeg dolazi kada se koriste N-grami višeg reda. Najverovatniji uzrok ove pojave je upravo visok procenat OOV reči, što je izazvalo često korišćenje *back-off* procedure, koja se često pokazuje kao loše rešenje (Mikolov, 2012).

Važno je još jednom napomenuti da su morfološke klase u okviru ovog eksperimenta definisane morfološkom anotacijom bez analize konteksta. Uz analizu konteksta može se očekivati da rezultati za morfološke modele budu značajno bolji. Ipak, najveća prednost morfoloških modela jeste mogućnost delimičnog prevazilaženja problema OOV reči. Naime, bez obzira na to da li se analiza konteksta vrši ili ne, morfološka anotacija može se izvesti ne samo na korpusu za obuku, već i na rečima koje su sadržane u morfološkom rečniku.

Drugim rečima, pri korišćenju morfoloških modela, za reči koje se nisu pojavile u korpusu za obuku moguće je odrediti morfološke klase ako su te reči sadržane u morfološkom rečniku. U tabeli 11 prikazani su rezultati WER za morfološke modele kojima su pridružene informacije o pripadanju reči iz morfološkog rečnika odgovarajućim morfološkim klasama. Vidi se da se time postiže smanjenje WER za sve funkcionalne stilove, što je najdrastičnije izraženo kod stilova kod kojih je prvobitan procenat OOV reči bio veoma visok (administrativni i naučni).

funkcionalni stil	bigram WER	trigram WER	kvadrigram WER
administrativni (C = 443)	24,66	25,13	25,24
literarni (C = 828)	27,51	27,78	28,58
naučni (C = 646)	22,44	23,00	23,30
novinarski (C = 836)	31,37	32,06	32,57

Tabela 11. Rezultati evaluacije morfoloških modela računanjem WER, u slučaju kada se koriste informacije o morfološkim klasama kojima pripadaju reči koje su sadržane u morfološkom rečniku (C - veličina rečnika)

Redukcija veličine jezičkih modela

Za potrebe aplikacija za uređaje sa relativno malim memorijskim kapacitetom ili relativno malom procesorskom snagom, neophodno je prilagoditi veličinu jezičkog modela. Takođe, kada se radi o sistemima za prepoznavanje govora, često je potrebno obučiti jezički model za unapred zadati rečnik, nasuprot obučavanju modela za rečnik koji se dobija na osnovu korpusa za obuku. Iz navedenih razloga potrebno je utvrditi na koji način se uz određenu redukciju veličine modela gubi najmanje korisnih informacija. Dve najčešće korišćene tehnike za redukciju veličine modela jezika implementirane su u okviru alata *SRILM*. Prva tehnika predstavlja prosto postavljanje praga za minimalni broj pojavljivanja N-grama. Druga tehnika zasniva se na iterativnom uklanjanju N-grama, čijim se izostavljanjem iz modela minimalno povećava perpleksnost,

odnosno dolazi do minimalnog porasta entropije. Ova tehnika podrazumeva postavljanje maksimalne vrednosti porasta perpleksnosti modela u jednoj iteraciji (Stolcke, 1998).

Za eksperiment za srpski jezik, koji je izvršen sa ciljem da se uporede dve pomenute tehnike, korišćen je kompletan tekstualni korpus, odnosno združeni korpusi za pojedine funkcionalne stilove (Ostrogonac et al, 2013). Za obuku modela korišćeno je 99% korpusa, dok je 1% korišćen za evaluaciju. Kreirani su *Katz* trigram modeli, i to dve grupe – jedna sa unapred definisanim rečnikom veličine 2.000 (1.770 osnovnih oblika reči – lema), a druga sa veličinom rečnika 10.000 (6.753 leme).

Rezultati za prvu tehniku prikazani su u tabeli 12. Evidentno je da se kod modela sa rečnikom veličine 2.000 većina *N*-grama pojavljuje veoma često (čak i više od 9 puta), te je čak i smanjenje modela na svega 20% početne veličine rezultovalo relativno malim povećanjem perpleksnosti, odnosno malim smanjenjem koeficijenta diskriminacije. Kod modela sa rečnikom veličine 10.000 već se vidi značajniji porast perpleksnosti sa povećanjem praga minimalnog broja pojavljivanja *N*-grama. Iz ovog razloga, modeli sa većim rečnicima korišćeni su za eksperiment vezan za drugu tehniku redukcije veličine. Rezultati tog eksperimenta prikazani su u tabeli 13. Ako se uporedi model iz tabele 12 čija je perpleksnost 160,1 sa modelom iz tabele 13 sa istom vrednošću perpleksnosti, vidi se da se tehnikom minimalnog porasta entropije dobija model istog kvaliteta kakav bi se dobio tehnikom povećanja minimalnog broja pojavljivanja N-grama, ali je model dobijen tehnikom minimalnog porasta entropije za preko 35% manjeg obima. U tabelama 14 i 15 dati su rezultati eksperimenata sa modelima baziranim na lemama. I u ovom slučaju tehnika minimalnog porasta entropije pokazala se kao adekvatnija, uz sličan odnos veličina modela kao i u eksperimentima sa modelima reči.

min. broj pojavljivanja *N*-grama	rečnik veličine 2.000			rečnik veličine 10.000		
	broj parametara (x10^3)	*ppl*	*KD*	broj parametara (x10^3)	*ppl*	*KD*
1	410	58,0	5,5	1.580	129,0	10,9
5	134	59,8	4,9	358	148,1	8,0
6	114	60,2	4,8	297	151,9	7,6
7	100	60,5	4,7	254	155,0	7,3
8	89	60,8	4,6	223	157,7	7,0
9	81	61,1	4,6	**200**	**160,1**	**6,8**

Tabela 12. Veličine modela i vrednosti perpleksnosti za tehniku redukcije postavljanjem minimalnog broja pojavljivanja *N*-grama

maksimalni porast perpleksnosti po iteraciji (x10^{-7})	Perpleksnost modela	broj parametara (x10^3)
1,0	130,0	1000
10,0	145,0	280
27,0	158,4	138
29,9	**160,1**	**128**
40,0	165,0	100
50,0	169,0	88
100,0	186,0	53

Tabela 13. Vrednosti perpleksnosti i veličine modela za tehniku redukcije minimalnim porastom entropije

min. broj pojavljivanja *N*-grama	rečnik veličine 1.170			rečnik veličine 6.753		
	broj parametara (x10^3)	*ppl*	*KD*	broj parametara (x10^3)	*ppl*	*KD*
1	824	69,0	6,7	2.906	121,0	15,3
5	266	72,0	5,7	680	142,0	10,6
6	223	72,8	5,6	556	146,7	10,1
7	193	73,5	5,4	470	150,9	9,5
8	170	74,1	5,3	407	154,4	9,1
9	153	74,6	5,2	**360**	**157,5**	**8,7**

Tabela 14. Veličine modela baziranih na lemama i vrednosti perpleksnosti za tehniku redukcije postavljanjem minimalnog broja pojavljivanja *N*-grama

maksimalni porast perpleksnosti po iteraciji (x10^{-7})	Perpleksnost modela	broj parametara (x10^3)
1	125,1	1396
10	150,9	295
13	155,8	241
14	**157,3**	**228**
15	158,6	216
30	174,9	124
100	215,4	46

Tabela 15. Vrednosti perpleksnosti i veličine modela baziranih na lemama, za tehniku redukcije minimalnim porastom entropije

Ovo istraživanje, zajedno sa ranije opisanim istraživanjima u okviru ovog odeljka, rezultovalo je stvaranjem baze znanja, odnosno skupa uopštenih pravila, na osnovu kojih se mogu efikasno kreirati adekvatni modeli jezika za različite domene primene automatskog prepoznavanja govora za srpski jezik.

5.2 Sinteza govora na osnovu teksta

Za sintezu govora na osnovu teksta jezički modeli ne predstavljaju komponentu od podjednako velikog značaja kao u slučaju automatskog prepoznavanja govora, makar ne u formama razmatranim u okviru ovog istraživanja. Međutim, okviru sistema za sintezu govora prvi korak je normalizacija teksta, koja podrazumeva, između ostalog, podelu teksta na rečenice (eng. *sentence boundary disambiguation* – *SBD*), koja se u velikoj meri oslanja na jezičke modele. Za srpski jezik, bez obzira na to da li se radi o konkatenativnoj sintezi (Sečujski et al, 2007), sintezi baziranoj na skrivenim Markovljevim modelima (Pakoci, 2012), ili sintezi baziranoj na neuronskim mrežama (Delić et al, 2017), svi postojeći sistemi se oslanjaju na modul za pretprocesiranje u okviru kog se normalizacija teksta vrši pomoću manuelno definisanih pravila. Ova pravila podrazumevaju analizu konteksta u vidu delova teksta u okolini simbola koji mogu predstavljati kraj rečenice („.“, „!“, „?“, „;“, „…“). U okviru alata *SRILM* implementirana je mogućnost podele teksta na rečenice pomoću jezičkog modela. S obzirom na to da su korpusi za obuku modela za srpski jezik velikim delom bili inicijalno podeljeni na rečenice, ili manuelno pregledani nakon podele koja je izvršena pomoću pomenutog sistema zasnovanog na pravilima, kao i zbog drugih faktora o kojima će biti reči u nastavku ovog odeljka, moglo se pretpostaviti da se pomoću jezičkih modela može doći do poboljšanja tačnosti u podeli teksta na rečenice.

Tačnost utvrđivanja lokacija krajeva rečenica u tekstovima iznosi oko 95% kada se prosto svaki od simbola koji mogu predstavljati kraj rečenice tretira kao da to i jeste. Korišćenjem pravila, koja su u okviru sistema za sintezu govora implementirana za srpski jezik, ova tačnost može se podići na oko 99.6% (ali, naravno, varira u zavisnosti od uzorka koji se koristi za testiranje). U (Read et al, 2012) analizirana je potreba za daljim istraživanjima na temu SBD. Naime, iako je moguće postići veoma visoku tačnost podele teksta na rečenice korišćenjem manuelno definisanih pravila, ukoliko se negde ipak dogodi greška, ona se propagira kroz kompletan sistem za sintezu govora i utiče na izgovor većeg broja fonema i reči, zbog čega je važno otkloniti ovakve greške u što većoj meri. Evaluacija tačnosti podele teksta na rečenice vezana je u izvesnoj meri za skup podataka koji se koristi za testiranje. Jedan od najvećih problema jeste činjenica da većina aplikacija podrazumeva da ulazni tekst odgovara spontanoj govornoj komunikaciji (razni web sadržaji). Drugi veliki problem su skraćenice, koje u novije vreme nastaju često u okviru spontane komunikacije i, iako za sada ne predstavljaju deo standardnog jezika, pojavljuju se učestalo u tekstovima i unose greške u SBD. Pored ovih problema, postoje i mnogi drugi, poput činjenice da se i sami simboli kojima se označavaju krajevi rečenica mogu pojaviti u različitim oblicima kada je, na primer, reč o tzv. *rich text* formatu. Sve pomenuto ukazuje na jezički model kao adekvatan alat za SBD.

Kada su u pitanju istraživanja koja su vršena za druge jezike, SBD tehnike mogu se svrstati u jednu od tri grupe: tehnike bazirane na pravilima, tehnike bazirane na mašinskom učenju koje podrazumevaju nadgledanu obuku i postojanje anotiranih korpusa i tehnike bazirane na mašinskom učenju koje podrazumevaju nenadgledanu obuku (i koje, iako ne zahtevaju anotirane korpuse, ipak podrazumevaju određeno empirijsko znanje koje je potrebno da bi sistem funkcionisao). Najviše istraživanja u oblasti SBD rađeno je za engleski jezik, iako su SBD sistemi uglavnom realizovani sa namerom da se mogu koristiti i za druge jezike, uz eventualnu potrebu za postojanjem anotiranog korpusa za svaki nov jezik za koji se implementira podrška. Najpoznatiji višejezični SBD sistemi su *iSentenizer-μ* (Wong et al, 2014), *Punkt* (Kiss & Strunk, 2006) i *MaxEnt* (Agarwal et al, 2005).

Nenadgledane tehnike mašinskog učenja uglavnom se, kao što je ranije napomenuto, oslanjaju na skup empirijski utvrđenih relevantnih podataka na osnovu kojih se iz neanotiranog korpusa može izvesti izdvajanje obeležja i obučavanje sistema. *Punkt* je primer ovakvog sistema. On je baziran na pretpostavci da je, kod većine jezika, za rešavanje problema određivanja granica rečenica najvažnija identifikacija skraćenica u tekstovima. Kriterijumi za identifikaciju skraćenica su empirijski ustanovljeni analizom velikog tekstualnog korpusa, odnosno analizom tipova reči koje se nalaze pre skraćenica u tekstovima. Podaci o lokaciji i tipovima reči koje prethode skraćenicama korišćeni su kao obeležja na osnovu kojih je *Punkt* obučen.

Nadgledane tehnike mašinskog učenja obično se baziraju na klasifikacionim i regresionim stablima (eng. *classification and regression trees – CARTs*), neuronskim mrežama (Romportl et al, 2003), kriterijumu maksimalne entropije. Primer sistema zasnovanog na CART je *iSentenizer-μ*, koji je baziran na algoritmu koji podrazumeva inkrementalno učenje, što ga čini pogodnim za adaptaciju sistema na nove domene primene. Neki od sistema baziranih na kriterijumu maksimalne entropije opisani su u (Reynar & Ratnaparkhi, 1997) i (Le et al, 2008), od kojih je prvi implementiran za više jezika, dok je drugi obučen na vijetnamskim tekstovima. Ovim sistemima nisu potrebe dodatne informacije, ali postoje i sistemi kao što je, na primer, *Satz* (Palmer & Hearst, 1997), za koje je pre obuke potrebno izvršiti određivanje vrsta reči koje čine korpus. Jedna od popularnih tehnika nadgledanog mašinskog učenja je i učenje bazirano na transformacijama (eng. *transformation-based learning – TBL*). Cilj ove tehnike je automatska ekstrakcija pravila na osnovu anotiranog korpusa (Stamatatos et al, 1999), mada je radi smanjenja vremena obuke često potrebno posedovati informacije o području primene sistema koji se obučava pomoću ove tehnike.

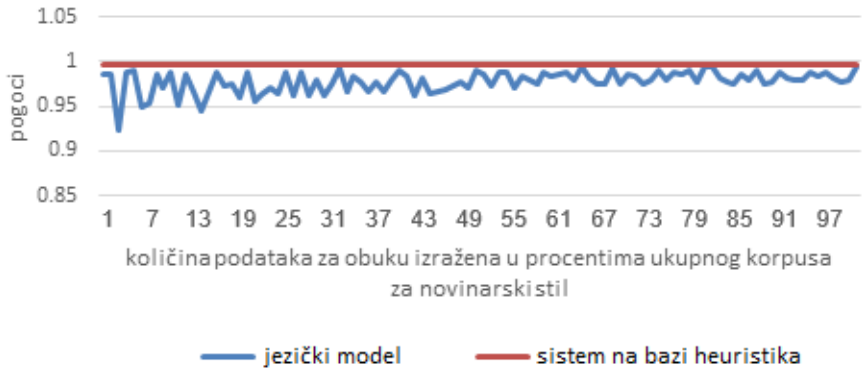
Hibridni pristup segmentaciji teksta na rečenice realizovan je za urdu (Rehman & Anwar, 2012). U okviru ovog istraživanja korišćen je unigram jezički model u kombinaciji sa manuelno implementiranim pravilima. Rezultati su pokazali izvesnu komplementarnost pomenuta dva pristupa. Ovi rezultati potvrđeni su u još jednom istraživanju čiji je cilj bilo poređenje sistema baziranih na pravilima i sistema baziranih na tehnikama mašinskog učenja (Wang & Huang, 2003). Međutim, u literaturi se korišćenje modela višeg reda u ove svrhe nigde ne pominje.

S obzirom na prirodu problema segmentacije teksta na rečenice, opravdano je bilo pretpostaviti da bi korpusi koji su oformljeni za srpski jezik (naravno, verzije koje sadrže znake interpunkcije, za razliku od verzija koje se koriste u prepoznavanju govora) mogli biti dovoljni za obuku modela koji bi se koristili za rešavanje ovog problema. Sa ciljem da se uporedi postojeći sistem baziran na pravilima sa jezičkim modelima u funkciji SBD sistema, izvršen je niz eksperimenata.

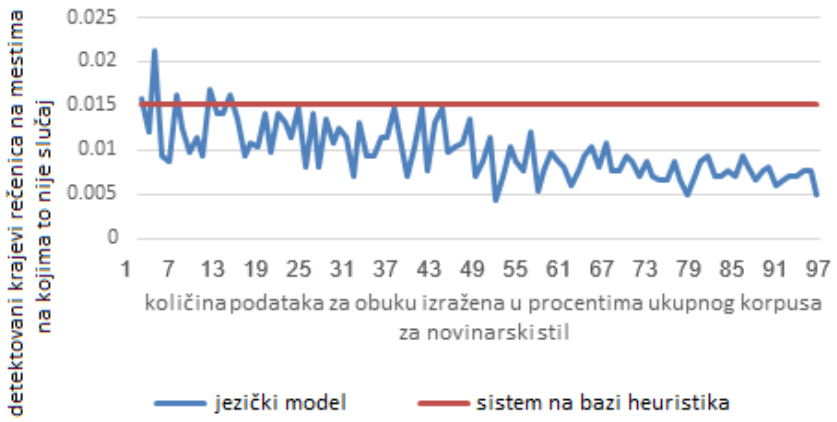
Za prvi eksperiment korišćen je novinarski korpus za obuku jezičkih modela, iz koga je izdvojen tekst za evaluaciju, koji je sadržao 1843 rečenice. Tekst za evaluaciju manuelno je pregledan kako bi se utvrdila tačnost podele na rečenice. Eksperiment je rađen sa bigram i trigram modelima i to tako što je deo korpusa koji je korišćen za obuku modela iterativno povećavan za po 1%, kako bi se utvrdilo kako količina podataka utiče na kvalitet modela, kada je u pitanju primena ovih modela u SBD. Vrednosti koje su od značaja za poređenje sa sistemom koji je baziran na pravilima (heuristikama) su pogoci (eng. *hits*) i promašaji, koji predstavljaju situacije u kojima jezički model postavlja oznaku za kraj rečenice, iako se zapravo ne radi o kraju rečenice (eng. *false positives*). Rezultati ovog eksperimenta prikazani su na slikama 20 i 21 za trigram modele (za bigram modele dobijeni su neznatno lošiji rezultati, ali se iz njih mogu izvesti isti zaključci). Vidi se da se pomoću jezičkog modela može postići isti procenat pogodaka kao kod sistema baziranog na pravilima, pri čemu je broj promašaja koje pravi jezički model mnogo manji. Takođe, vidi se da se dobri rezultati mogu postići i sa relativno malim korpusom za obuku. Skokovi i padovi koji se mogu uočiti kod krivih koje se odnose na rezultate jezičkog modela govore o tome da postoje delovi korpusa koji nepovoljno utiču na performanse modela kada je u pitanju zadatak SBD.

Upravo zbog rezultata prvog eksperimenta, na osnovu kojih se moglo zaključiti da je i mnogo manji korpus od postojećeg dovoljan za obučavanje jezičkih modela za SBD pod uslovom da se izabere pogodan podskup, izvršen je i eksperiment čiji je cilj bilo ispitivanje uticaja potkresivanja na performanse modela u okviru primene u segmentaciji teksta na rečenice. Za ovaj eksperiment korišćen

je model koji je u okviru prvog eksperimenta obučen na celokupnom korpusu za obuku za novinarski stil. Za potkresivanje je korišćena tehnika minimalnog porasta entropije, o kojoj je bilo reči u odeljku 5.1. U tabeli 16 dati su rezultati ovog ispitivanja.



Slika 20. Normalizovane vrednosti stope ispravno određenih lokacija krajeva rečenica pomoću trigram modela jezika i pomoću sistema zasnovanog na heuristikama (pravilima)



Slika 21. Normalizovane vrednosti stope detektovanih krajeva rečenica na mestima gde to nije slučaj, pomoću trigram modela jezika i pomoću sistema zasnovanog na heuristikama (pravilima)

koeficijent potkresivanja ($x10^{-7}$)	zadržano unigrama	zadržano bigrama	zadržano trigrama	pogodaka (na 1.843 rečenice)	Promašaja (na 1.843 rečenice)
0	365.585	4.086.488	2.016.314	1.833	9
1	365.585	1.898.933	572.904	1.812	10
5	365.585	501.690	146.629	1.809	23
10	365.585	254.619	66.329	1.795	32
50	365.585	55.751	8.478	1.829	60
100	365.585	25.198	3.308	1.826	85
500	365.585	2.704	320	1.834	168
1.000	365.585	905	80	1.832	178
5.000	365.585	51	5	1.836	212
10.000	365.585	16	2	1.836	259

Tabela 16. Uticaj potkresivanja na performanse jezičkog modela u segmentaciji teksta na rečenice

Kao što se može primetiti, broj pogodaka gotovo da nema mnogo veze sa veličinom modela, odnosno i vrlo mali modeli mogu da detektuju krajeve rečenica na mestima gde oni zaista i postoje. Međutim, potkresivanje modela povećava učestalost detekcije kraja rečenice na mestima gde to nije slučaj. Fluktuacije broja pogodaka upućuju na potrebu za detekcijom i izuzimanjem iz obuke delova korpusa koji evidentno negativno utiču na tačnost segmentacije.

Ono što je ostalo da se ispita jeste ponašanje modela koji su obučeni za određeni funkcionalni stil, kada se primene na tekstu koji pripada nekom drugom stilu. Ispitani su modeli obučeni na korpusu za novinarski, literarni stil, kao i modeli koji su obučeni na združenim korpusima svih stilova. U tabeli 17 prikazani su rezultati ovog eksperimenta. Važno je napomenuti da je test skup koji je predstavljao novinarski stil sadržao 1.843 rečenice (isti skup koji je korišćen u prethodna dva eksperimenta), dok je test skup koji je predstavljao literarni stil sadržao 3.023 rečenice, a test skup koji je predstavljao združene korpuse 2.153 rečenice. U skladu sa ovim podacima treba tumačiti i dobijene rezultate. Kao što se može zaključiti, funkcionalni stil je od izuzetno velikog značaja, kada se jezički modeli kreiraju sa ciljem da se primenjuju u segmentaciji teksta na rečenice. Modeli obučeni na združenim korpusima postigli su odlične rezultate, a zanimljivo je da su ovi modeli bili čak uspešniji od modela koji su obučavani na korpusu za

literarni stil, kada su upoređene njihove performanse na test skupu koji predstavlja literarni stil.

obučen na/primenjen na	bigram		trigram	
	pogoci	promašaji	pogoci	promašaji
novinarski/novinarski	**1.830**	**14**	**1.833**	**9**
novinarski/literarni	2.278	306	2.298	316
novinarski /celokupni	1.927	94	1.918	96
literarni/novinarski	1.375	180	1.341	119
literarni/literarni	**2.571**	**247**	**2.645**	**238**
literarni/celokupni	1.768	184	1.736	124
celokupni/novinarski	1.832	23	1.820	22
celokupni/literarni	2.641	188	2.802	191
celokupni/celokupni	**2.039**	**66**	**2.036**	**66**

Tabela 17. Performase modela u segmentaciji teksta kada su obučavani na korpusu za određeni stil, a primenjeni na tekstu koji pripada istom ili različitom stilu

5.3 Automatska detekcija i korekcija grešaka u tekstovima

U oblasti obrade prirodnog jezika, detekcija grešaka u tekstovima predstavlja jedan od velikih izazova. Aplikacije za detekciju i korekciju grešaka u tekstovima razvijene su za mnoge jezike, ali sa različitim nivoima tačnosti i uopšte, mogućnosti u smislu tipova grešaka koje one mogu da otkriju. Postoje tri vrste grešaka koje bi savremeni sistem za detekciju i korekciju grešaka trebalo da tretira.

U prvu grupu spadaju greške koje mogu nastati na različite načine, ali manifestuju se kao reči koje ne postoje u rečniku (eng. *non-word errors*). Za engleski jezik, ove greške su klasifikovane prema uzroku nastanka na tipografske, kognitivne i fonetske (Liang, 2005). Na srpski jezik, međutim, ova podela nije primenjiva, jer je fonetski zapis reči uglavnom u skladu sa izgovorom. Ovakve greške su najlakše za detekciju i korekciju. Naime, detekcija ovakvih grešaka moguća je prostom proverom postojanja reči u rečniku. Za izvršenje ovog procesa, neophodno je, naravno, izdvojiti reči iz teksta (odvojiti znake interpunkcije, izuzeti iz pretrage određene konstrukcije poput e-mail adresa ili adresa web sajtova itd.), a tačnost detekcije ovakvih grešaka zavisi od obima rečnika kojim se raspolaže za

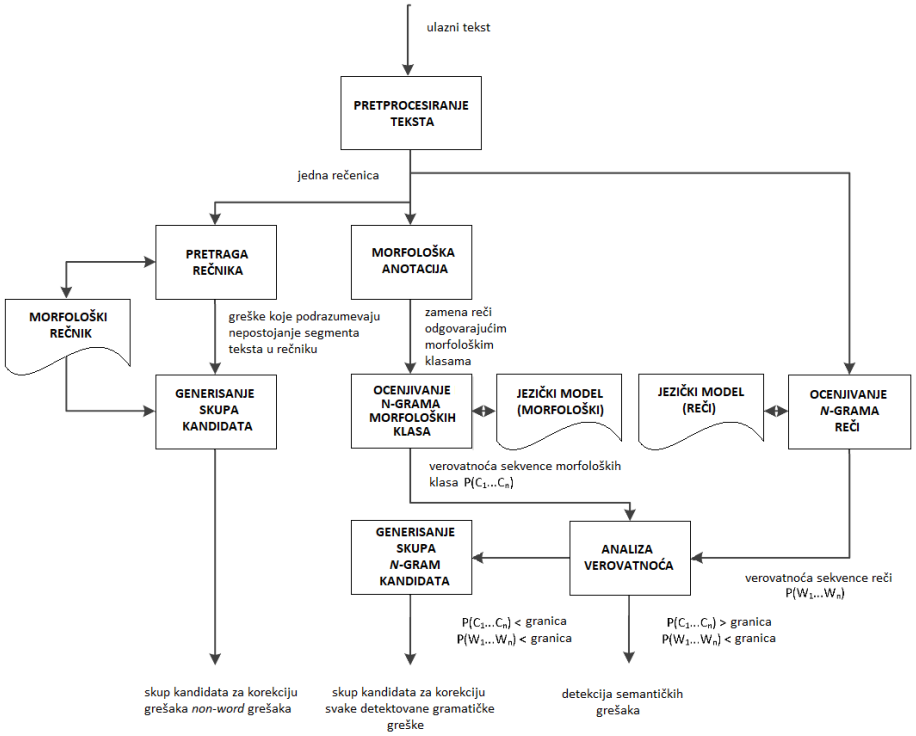
određeni jezik. Kada je u pitanju provera postojanja reči u rečniku, veoma je važno da takva pretraga bude što efikasnija, kako bi proces detekcije i korekcije grešaka mogao da se odvija u realnom vremenu, odnosno u toku pisanja tekstualnog sadržaja. U ovu svrhu su razvijene mnoge tehnike pretrage (De Schryver & Prinsloo, 2004). Što se tiče korekcije grešaka, princip je isti kao i za druge tipove grešaka. Sistem pronalazi listu reči – kandidata, obično na osnovu sličnosti fonetske transkripcije (Navarro, 2001), a korisnik iz te liste odabira reč kojom želi da koriguje grešku. Ukoliko je cilj automatska korekcija greške, onda se umesto liste bira jedna reč – najpogodniji kandidat, i tom rečju se zamenjuje pogrešno napisan segment teksta. Kao kod same pretrage rečnika, kod procesa nalaženja kandidata za korekciju greške takođe je potrebno obezbediti maksimalnu efikasnost. Za ovo se najčešće koriste određene heuristike (Mozgovoy, 2011). Primer je zamena mesta dva susedna slova, ili zamena slova slovom za koje se odgovarajući taster na standardnim tastaturama nalazi u neposrednoj blizini tastera za slovo koje se zamenjuje.

Preostale dve vrste grešaka su gramatičke i semantičke greške. Ove vrste grešaka zahtevaju poseban tretman (Verberne, 2002), koji podrazumeva analizu konteksta. Za analizu konteksta je neophodno imati jezički model u nekoj formi, kao i način izdvajanja morfoloških informacija (Vosse, 1992), kako bi se ovakve greške mogle detektovati sa prihvatljivom tačnošću. Korekcija gramatičkih i semantičkih grešaka je takođe komplikovanija od korekcije prve grupe grešaka, s obzirom na to da reč kojom bi se trebala ispraviti greška koja pripada nekoj od pomenutih grupa ne mora biti slična originalnom segmentu teksta po fonetskoj transkripciji. Za gramatičke greške, ovo zavisi, pre svega, od pravila konstruisanja izvedenih oblika reči, a ta pravila variraju značajno od jezika do jezika. Semantičke greške, ipak, predstavljaju najveći izazov, jer je potrebno da aplikacija za detekciju i korekciju grešaka na neki način izdvoji informacije o značenjima pojedinih reči iz korpusa za obuku.

Važno je napomenuti da je klasifikacija grešaka prema pomenutim trima grupama definisana na opisan način iz praktičnih razloga, zbog grupisanja tehnika za tretiranje pojedinih vrsta grešaka. U realnosti, situacija je mnogo komplikovanija.

Tokom razvoja govornih tehnologija na Fakultetu tehničkih nauka, razvijena je i aplikacija pod nazivom *anSpellChecker*, koja se oslanjala na rečnik radi detekcije grešaka u tekstovima (Bojanić et al, 2012). Ova aplikacija, ipak, nije pružala mogućnost detekcije i korekcije gramatičkih i semantičkih grešaka. Sličnu aplikaciju za srpski jezik razvila je i kompanija *Microsoft*, u vidu dodatka osnovnom alatu *Microsoft Word*. U okviru ovog istraživanja, prvi koraci ka kreiranju aplikacije

za detekciju i korekciju gramatičkih i semantičkih grešaka opisani su u (Ostrogonac et al, 2015a). Na slici 22 prikazana je struktura ovakve aplikacije.



Slika 22. Šematski prikaz arhitekture sistema za detekciju i korekciju gramatičkih i semantičkih grešaka u tekstovima na srpskom jeziku

Ideja je da se paralelno analiziraju izlazi osnovnog jezičkog modela – modela reči i jezičkog modela baziranog na morfološkim klasama, pri čemu se kod morfološkog modela posmatraju upravo verovatnoće sekvenci klasa (te verovatnoće se, dakle, ne množe verovatnoćama da pojedine reči pripadaju određenim klasama). Greške koje rezultuju segmentima teksta koji ne postoje u rečniku detektuju se prostom pretragom među pojavnim oblicima reči koje su sadržane u morfološkom rečniku, a za korekciju se kandidati traže pomoću algoritma minimalog rastojanja uređivanja (eng. *Minimal Edit Distance*) (Levenshtein, 1966). Za gramatičke greške se očekuje da verovatnoća *N*-grama bude mala i na izlazu modela reči i na izlazu morfološkog modela, dok bi semantičke greške mogle da se manifestuju malom verovatnoćom na izlazu modela reči, dok bi morfološki model dao verovatnoću uobičajenu za „ispravne“ sekvence morfoloških klasa.

Cilj preliminarnog ispitivanja mogućnosti korišćenja morfoloških modela za detekciju netrivijalnih tipova grešaka u sklopu strukture koja je prikazana na slici 22 bio je sa se utvrdi da li se na nivou rečenice mogu uočiti razlike u verovatnoćama koje daju model reči i morfološki model pre i nakon što se jedna ili više reči u autentičnom tekstu zameni rečju kojom se izaziva gramatička ili semantička greška (Ostrogonac et al, 2015a). Rezultati ovog ispitivanja bili su ohrabrujući, a prikazani su u tabelama 18 i 19, za trigram modele obučene na korpusu za novinarski stil. U tabeli 18 vidi se ponašanje modela reči i morfološkog modela u situacijama kada se u rečenicama pojave semantičke greške. U 7 od 10 slučajeva, model reči je rečenicama sa greškom dodelio značajno niže verovatnoće. S druge strane, morfološki model se ponašao kao što je bilo očekivano – s obzirom na to da je većina ovih rečenica gramatički korektna, verovatnoće rečenica sa semantičkim greškama koje su se dobile na izlazu modela bile su iste ili vrlo slične verovatnoćama ispravnih rečenica. Ovde je važno napomenuti da jezički modeli koji su korišćeni u okviru ovih eksperimenata, ne prave razliku između velikih i malih slova. To može objasniti situaciju zašto je u primeru *Novak Đoković je pobedio* ispalo verovatnije da naredna reč bude *krušku*, nego *Nadala*, jer se ova lična imenica u korpusu za obuku nije razlikovala od glagola sa istom transkripcijom. Može se, dakle, očekivati da se pravljenjem razlike između malih i velikih slova dodatno poveća tačnost modela u razlikovanju ispravnih rečenica i rečenica sa semantičkim greškama. U tabeli 19 prikazani su rezultati za situacije kada se u rečenicama pojavljuju gramatičke (sintaksne) greške. Ove greške je, evidentno, znatno lakše detektovati nego semantičke greške. Morfološki modeli su dodelili manju verovatnoću rečenicama sa greškom u 9 od 10 slučajeva, a u jednom slučaju je rečenici sa greškom dodeljena ista verovatnoću kao i ispravnoj rečenici. Model reči je takođe ispravno reagovao na pojavljivanje gramatičkih grešaka u 9 od 10 slučajeva.

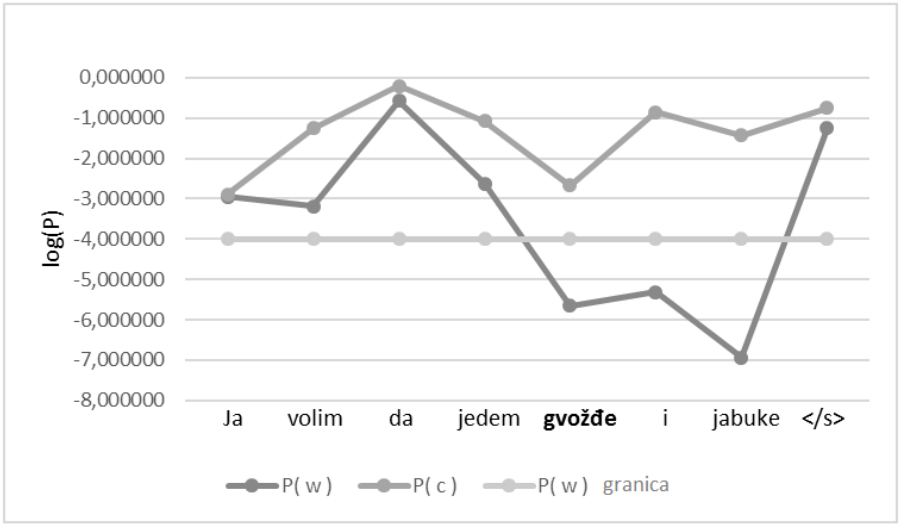
Nakon preliminarnog ispitivanja bilo je potrebno utvrditi kako se detektuje tačna lokacija greške u rečenici, kao i načine na koje je moguće utvrditi granice za odlučivanje o tome da li se u nekoj konkretnoj situaciji radi o grešci ili ne. U narednom istraživanju došlo se do zaključaka o mogućim načinima rešavanja ovih problema (Ostrogonac, 2016). Na slici 23 prikazana je analiza izlaza modela reči i morfološkog modela, korak po korak, za sekvencu u okviru koje se pojavljuje semantička greška. U zavisnosti od reda *N*-gram modela, vidi se da će pad izlazne verovatnoće modela reči da se propagira na nekoliko koraka, a lokacija greške je onda na prvom koraku kod koga dolazi do pada verovatnoće. Međutim, tu se može javiti problem ukoliko dođe do spuštanja na niži nivo *N*-grama (*back-off*). U tom slučaju, mogu se dogoditi različite situacije, koje su ilustrovane na slici 24, za trigram model, dok je na slici 25 prikazano nekoliko primera realnih situacija.

Rečenice: Autentične (ispravne) Sa jednom ili dve greške	izlaz modela reči	izlaz morfološkog modela
1A: *Đoković je pobedio Nadala.*	-8,37	-9,22
1B: *Đoković je pobedio krušku.*	**-6,55**	**-5,31**
2A: *Imam psa i mačku.*	-17,87	-8,69
2B: *Imam psa i tačku.*	**-17,29**	**-8,69**
3A: *Dođi ovde gde smo svi.*	-20,69	-12,85
3B: *Dođi ovde nigde smo svi.*	-24,86	-13,04
4A: *Uradi domaći zadatak!*	-12,45	-5,65
4B: *Uradi domaći telefon!*	-16,18	-5,73
5A: *Sanjao sam san.*	-10,69	-6,28
5B: *Sašio sam san.*	**-8,37**	**-6,28**
6A: *Radim po ceo dan i noć.*	-16,20	-9,85
6B: *Radim po ceo toranj i noć.*	-25,38	-9,85
7A: *Cela ulica je poplavljena.*	-17,07	-5,37
7B: *Cela mačka je poplavljena.*	-20,17	-5,37
8A: *Jabuka je zrela.*	-15,28	-4,64
8B: *Jabuka je popravljena.*	-15,96	-4,64
9A: *Beba nosi pelene.*	-17,45	-6,39
9B: Beda nosi pelene.	-19,39	-6,30
10A: *Nisam to ni slutio.*	-13,56	-12,73
10B: *Jesam to ni slutio.*	-15,95	-13,72

Tabela 18. Logaritmi verovatnoća – izlazi modela reči i morfološkog modela za parove rečenica od kojih se svaki sastoji od ispravne rečenice i rečenice koja sadrži semantičku grešku

Rečenice: Autentične (ispravne) Sa jednom ili dve greške	izlaz modela reči	izlaz morfološkog modela
1A: *On je skuvao čaj.*	-15,52	-5.77
1B: *On je skuvala čaj.*	-16,81	-8,77
2A: *Snežana tečno govori srpski.*	-18,04	-11,15
2B: *Snežana tečnim govori srpski.*	-20,86	-12,93
3A: *Žitelji Beograda su ogorčeni.*	-16,64	-7,28
3B: *Žitelji beogradskog su ogorčeni.*	-17,67	-10,06
4A: *Nemanja je uradio domaći.*	-15,32	-10,47
4B: *Nemanja će uradio domaći.*	-17,61	-12,28
5A: *Došao je u devet sati.*	-10,68	-6,34
5B: *Došao je u devetih sati.*	-17,06	-9,72
6A: *Izvučeno je sedam brojeva.*	-17,75	-5,78
6B: *Izvučeno smo sedam brojeva.*	-19,83	-7,74
7A: *Ovo je grad budućnosti.*	-10,75	-6,22
7B: *Ovoga je gradovi budućnosti.*	-18,40	-12,65
8A: *Živim u sedmoj zgradi u ulici.*	-19,21	-11,46
8B: *Živim u sedam zgradi u ulici.*	**-18,67**	-15,02
9A: *Kupićemo nove cipele.*	-16,22	-7,56
9B: *Kupićemo novoga cipele.*	-18,71	-11,80
10A: *Sada ćemo to i dokazati.*	-15,39	-10,01
10B: *Sada ćemo to i dokazavši.*	-16,85	**-10,01**

Tabela 19. Logaritmi verovatnoća – izlazi modela reči i morfološkog modela za parove rečenica od kojih se svaki sastoji od ispravne rečenice i rečenice koja sadrži gramatičku (sintaksnu) grešku

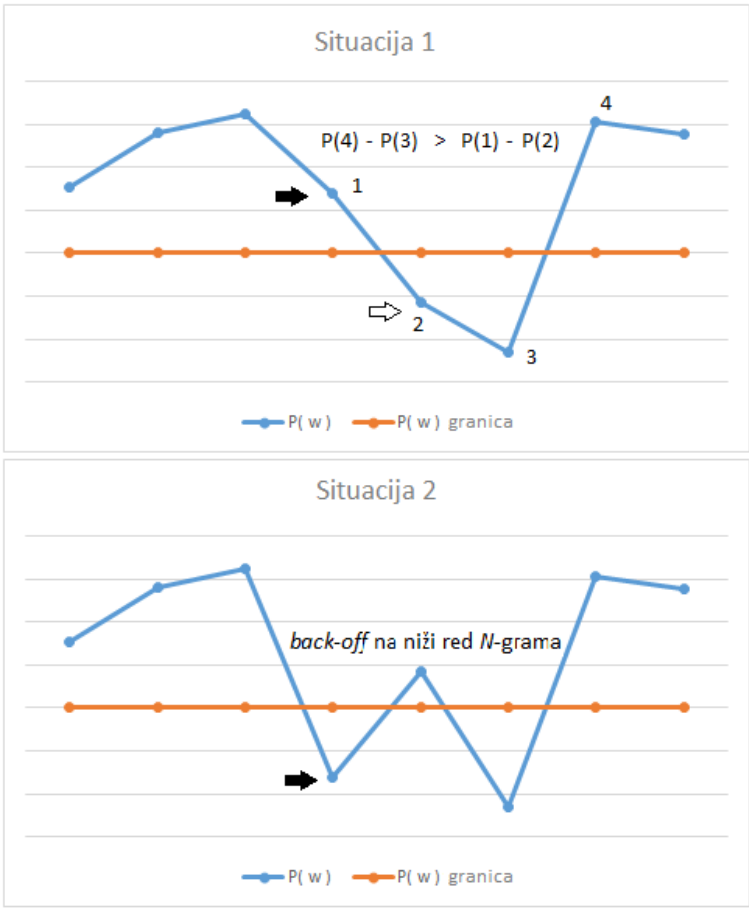


Slika 23. Primer verovatnoća *N*-grama koje se dobijaju pomoću modela reči i morfološkog modela u okviru sekvence koja sadrži semantičku grešku

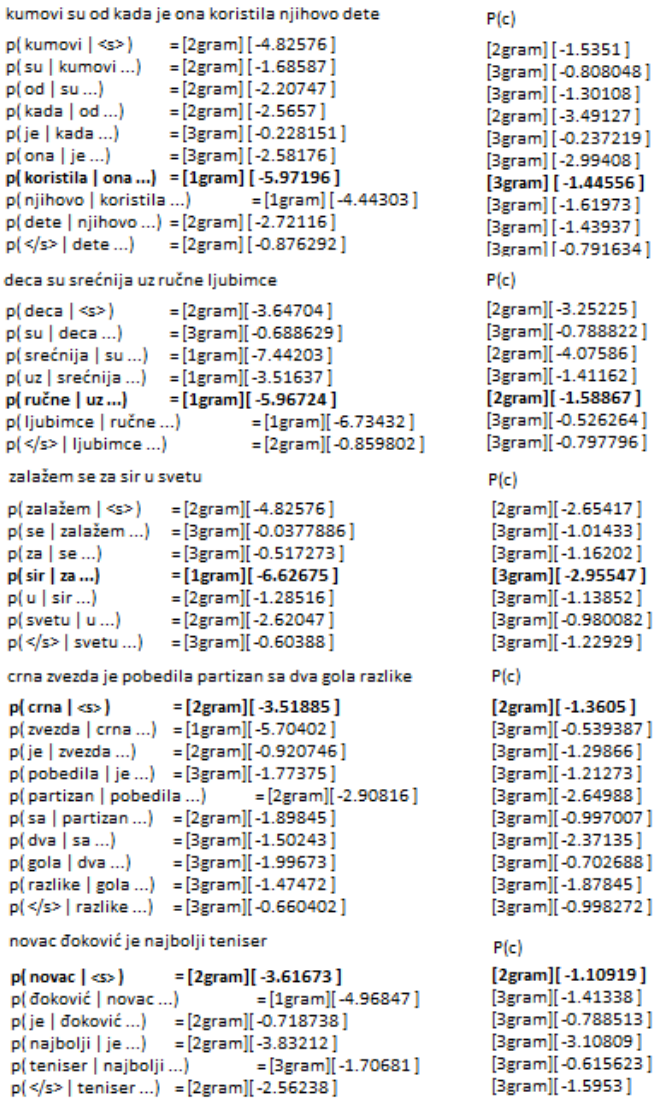
U prvoj situaciji, gde postoji pad verovatnoće u jednom koraku, ali je tek u naredna dva koraka taj pad značajan, odluka o lokaciji greške može se doneti poređenjem razlike verovatnoća u tačkama 1 i 2 sa razlikom verovatnoća u tačkama 3 i 4. Analiza sličnih situacija potvrđuje da, ukoliko je skok verovatnoće između tačaka 3 i 4 veći od pada verovatnoće između tačaka 1 i 2, lokacija greške je najverovatnije u tački 1. U suprotnom, lokacija greške je najverovatnije u tački 2. U drugoj situaciji sa slike 24, potrebno je analizirati do kog nivoa je vršena *back-off* procedura u tački koja se nalazi između dva koraka u kojima je verovatnoća izrazito mala. Ukoliko je došlo do *back-off* procedure, najverovatnije je da jeste u pitanju semantička greška i da je ona locirana u prvoj tački u kojoj je verovatnoća izrazito mala. Ukoliko se, ipak, izraziti pad verovatnoće uočava samo u jednoj tački, najverovatnije je da je u pitanju prosto *N*-gram koji se retko pojavljivao u korpusu za obuku, a ne semantička greška (osim ako je u pitanju poslednja reč, u kom slučaju je opet od koristi informacija o tome da li je došlo do *back-off* procedure). Ako se, primera radi, dogodi da se za trigram model javi pad verovatnoće u više od 3 tačke, najverovatnije je da postoji greška, ili više njih, u kombinaciji sa retkim *N*-gramima. U ovakvim situacijama nije moguće izvršiti preciznu analizu, pa je najbolje rešenje korisniku dati do znanja da bi trebalo da obrati pažnju na kritičnu sekvencu reči i da sam odluči o tome da li je negde došlo do greške.

Ono što svakako predstavlja najveći problem jeste utvrđivanje granica na osnovu kojih se odlučuje o tome da li se radi o greškama ili prosto o *N*-gramima koji su retko viđeni u korpusu za obuku. Ima više načina da se ove granice utvrde.

Moguće je koristiti prosečnu verovatnoću po reči, izračunatu na nivou rečenice (ili čak na širem kontekstu, ako je to moguće) i na osnovu nje odrediti granicu. Mnogo komplikovanija, ali verovatno i pogodnija tehnika, mogla bi se izvesti dodavanjem grešaka, bilo da su u pitanju gramatičke ili semantičke greške, u rečenice određenog korpusa i analizom rezultata, odnosno procene verovatnoće rečenica od strane jezičkog modela pre i posle uvođenja grešaka. Tako bi se na osnovu ovakve dve verzije korpusa mogao obučiti sistem, koji bi se oslanjao na bazu graničnih vrednosti i analizu konteksta, na osnovu koje bi odlučivao o tome koju granicu je pogodno primeniti u kojoj situaciji. Grupisanje reči na osnovu semantike bi takođe moglo biti od velike koristi za podešavanje sistema za detekciju i korekciju netrivijalnih tipova grešaka u tekstovima.



Slika 24. Primer verovatnoća *N*-grama koje se dobijaju pomoću modela reči i morfološkog modela u okviru sekvence koja sadrži semantičku grešku



Slika 25. Primeri verovatnoća *N*-grama koje se dobijaju pomoću trigram modela za sekvence koje sadrže semantičke greške

5.4 Ostale primene

Osim u tehnologijama o kojima je bilo reči u prethodnim odeljcima ovog poglavlja, jezički modeli igraju veoma važnu ulogu i u mnogim drugim

tehnologijama. Jedna od ovih tehnologija je mašinsko prevođenje, za koje se koriste obimni jezički modeli (Brants et al, 2007). Kompanija *Microsoft* je razvila sistem za mašinsko prevođenje tekstova sa srpskog na mnoge druge jezike. Poznato je, međutim, da tačnost ovog sistema nije ni blizu zadovoljavajuće i da on može poslužiti eventualno za prevođenje kratkih sekvenci, kada je cilj razumeti smisao poruke, a ne dobiti korektan prevod. Mašinsko prevođenje predstavlja jedan od najtežih istraživačkih izazova u oblasti obrade prirodnog jezika, i u ovoj oblasti i dalje postoji mnogo prostora za unapređivanje postojećih tehnika. Još jedna tehnologija u kojoj jezički modeli igraju važnu ulogu jeste kompresija tekstualnih sadržaja (El Daher & Connor, 2016). Kompresija tekstualnih sadržaja naročito je dobila na značaju otkad je počela nagla ekspanzija količine tekstualnih sadržaja koji se čuvaju u digitalnoj formi, čemu su doprinele društvene mreže, kao i elektronska komunikacija uopšte. Ekstrakcija informacija (eng. *Information Retrieval*) predstavlja još jednu oblast u kojoj jezički modeli igraju važnu ulogu (Sanderson & Bruce Croft, 2012). Za potrebe rekonstrukcije informacija razvijani su različiti jezički modeli, u zavisnosti od domena primene (Larson, 2001) (Lavrenko & Croft, 2001) (Song & Bruce Croft, 1999). Klasifikacija dokumenata po sadržaju korišćenjem jezičkih modela takođe predstavlja važnu oblast istraživanja (Majdoubi et al, 2010), (Ostrogonac et al, 2016).

Za srpski jezik postoji dobra osnova za razvoj ili usavršavanje svih pomenutih tehnologija, kao i mnogih drugih. U daljim istraživanjima težnja bi trebalo da bude usmerena ka automatskom izdvajanju semantičkih informacija iz tekstualnog sadržaja. Ideja za realizaciju sistema za grupisanje reči po semantičkoj sličnosti izložena je u (Ostrogonac et al, 2015b). Pored toga, neophodno je raditi na proširenju tekstualnih korpusa za obuku modela. U prilogu 5 dati su primeri rečenica koje su generisane od strane različitih modela koji su u okviru ovog istraživanja kreirani za srpski jezik. Evidentno je da postoji još mnogo prostora za poboljšanje kvaliteta modela. Međutim, može se očekivati da će se u narednim godinama razviti nove tehnike za rešavanje pojedinih problema iz oblasti obrade prirodnog jezika kombinacijom različitih modela i informacija koje se mogu dobiti analizom tekstualnih sadržaja. Daljim razvojem računara i porastom memorijskih i procesorskih kapaciteta mobilnih uređaja stvoriće se nove mogućnosti za korišćenje neuronskih mreža, a možda i nekih novih tehnika mašinskog učenja.

POGLAVLJE 6

ZAKLJUČAK

Jezički modeli predstavljaju važnu komponentu sistema koji se baziraju na govornim i jezičkim tehnologijama. Imajući to u vidu, važan doprinos istraživanja koje je opisano u ovoj disertaciji ogleda se u razvoju pomenutih tehnologija za srpski jezik.

Tokom protekle dve decenije, za srpski jezik je prikupljena i razvijena velika količina različitih govornih i jezičkih korpusa, što je omogućilo razvoj visokokvalitetnih sistema za automatsko prepoznavanje govora, sintezu govora na osnovu teksta, kao i detekciju štamparskih, pravopisnih i gramatičkih grešaka u tekstovima. Međutim, jezički modeli koji su korišćeni u pomenutim sistemima, predstavljaju skupove ručno implementiranih pravila i imali su ograničene mogućnosti. Predmet ovog istraživanja bio je razvoj savremenih jezičkih modela za srpski jezik, kako bi se potpomogao dalji razvoj govornih i jezičkih tehnologija. Prvi deo istraživanja obuhvatao je prikupljanje, klasifikaciju i obradu tekstualnih sadržaja na srpskom jeziku. Ovaj proces rezultovao je postojanjem tekstualnih korpusa koji ukupno sadrže preko 22 miliona reči. Ipak, za visoko inflektivan jezik kakav je srpski, za određene primene ovi korpusi nisu bili dovoljno obimni. Stoga je bilo potrebno kreirati klasne modele, sa ciljem da se oni, samostalno ili u kombinaciji sa osnovnim modelima, koriste radi postizanja boljih performansi u okviru sistema za koje su namenjeni. Grupisanje reči izvršeno je na osnovu morfoloških informacija, koje su dobijene iz tekstualnih sadržaja putem automatske morfološke anotacije, koja je implementirana tokom ranijih istraživanja (Sečujski, 2009). Eksperimenti su pokazali da je grupisanje na osnovu morfoloških informacija značajno pogodnije za modelovanje jezika nego što je to klasterizacija na osnovu statistika tekstualnog korpusa.

U okviru ovog istraživanja, dobijeni jezički modeli (osnovni i klasni) testirani su u različitim sistemima. Pokazano je povećanje tačnosti ASR sistema u vidu smanjenja verovatnoće greške prepoznavanja na nivou reči, korišćenjem klasnih jezičkih modela koji se baziraju na morfološkim informacijama u odnosu na klasne modele kod kojih su klase izvedene na osnovu statistika tekstualnog korpusa. Korišćenjem jezičkih modela radi podele teksta na rečenice u okviru pretprocesiranja teksta pri sintezi govora takođe je postignuto povećanje tačnosti. Kako bi se iskoristio potencijal morfoloških (klasnih) jezičkih modela u detekciji i korekciji gramatičkih i semantičkih grešaka, kreiran je prototip ovakvog sistema. Rezultati prvih eksperimenata pokazali su da je kombinacijom osnovnog i

morfološkog modela zaista moguće detektovati pomenute tipove grešaka u tekstu, što predstavlja velik iskorak u odnosu na postojeće sisteme, koji se uglavnom oslanjaju na rečnike, odnosno prostu pretragu radi detekcije grešaka.

Nastavak istraživanja prvenstveno će se kretati u pravcu finog podešavanja parametara modela za različite primene, kao i nalaženja najboljih kombinacija modela u zavisnosti od ograničenja konkretnih aplikacija. Pored toga, izdvajanje semantičkih informacija, odnosno grupisanje reči na osnovu semantike, naročito u vidu hijerarhijske strukture, predstavlja izuzetno značajan pravac daljeg istraživanja. Iako je krajnji cilj obrade prirodnog jezika zapravo razumevanje jezika od strane mašine (koje bi trebalo biti nalik na ljudsko) (Manning & Schütze, 1999), put do tog cilja za sada podrazumeva rešavanje pojedinačnih zadataka ove naučne oblasti, kako bi se prirodnost komunikacije između čoveka i mašine postepeno povećavala. Dalji napredak u ovoj oblasti od naročitog je značaja i u edukaciji. Programska podrška za edukaciju baziranu na govornim i jezičkim tehnologijama od posebne je vrednosti za slepe i slabovide učenike. Za potrebe edukacije slepih i slabovidih već su razvijene edukativne aplikacije, odnosno, igre bazirane na govornim tehnologijama za srpski jezik (Lučić et al, 2015). Ipak, u korišćenim ASR sistemima akustički modeli za sada nisu adaptirani na dečje glasove, te je uticaj jezičkih modela na tačnost prepoznavanja govora u ovim aplikacijama veoma važan.

Primena jezičkih modela u svim pomenutim tehnologijama, kao i njihova primena u obrazovanju, predstavlja važan faktor u očuvanju srpskog jezika, a to je, u opštem smislu, jedan od najvažnijih doprinosa ove disertacije.

LITERATURA

Agarwal N., Ford K. H., Shneider M., 2005. *Sentence Boundary Detection Using a MaxEnt Classifier*.

Arisoy E., Sainath T. N., Kingsbury B., Ramabhadran B., 2012. *Deep neural network language models*. Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, June 8, Montreal, Canada, pp. 20-28.

Bengio Y., Simard P., Frasconi P., 1994. *Learning Long-Term Dependencies with Gradient Descent is Difficult*. IEEE Transactions on Neural Networks, No. 5, pp. 157-166.

Bengio Y., Ducharme R., Vincent P., Jauvin C., 2003. *A Neural Probabilistic Language Model*. Journal of Machine Learning Research, vol. 3, pp. 1137-1155.

Blei D., Ng A., Jordan M., January 2003. *Latent Dirichlet Allocation*. Journal of Machine Learning Research vol. 3, No. 4–5, pp. 993-1022, DOI 10.1162/jmlr.2003.3.4-5.993.

Boden M., 2002. *A Guide to Recurrent Neural Networks and Backpropagation*. In the Dallas project, SICS Technical Report T2002:03.

Bojanić M., Ostrogonac S., Sečujski M., Vujnović-Sedlar N., Suzić S., 2012. *A detector of spelling errors for Serbian - anSpellChecker*. Technical Solution: Prototype, Faculty of Technical Sciences and AlfaNum, Novi Sad, Serbia.

Brants T., Popat A., Xu P., Och F., Dean J., 2007. *Large Language Models in Machine Translation*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, pp. 858-867.

Broman S., Kurrimo M., September 2005. *Methods for Combining Language Models in Speech Recognition*. Proceedings of 9th European Conference on Speech Communication and Technology, pp. 1317-1320.

Brown M. F., De Souza P. V., Mercer R. L., Della Pietra V. J., Lai J. C., December 1992. *Class-Based n-gram Models of Natural Language*. Computational Linguistics, vol. 18, No. 4, pp. 467-479.

Chen S., Beeferman D., Rosenfeld R., 1998. *Evaluation metrics for language models*. DARPA Broadcast News Transcription and Understanding Workshop, pp. 275-280.

Chen S., Goodman J., October 1999. *An empirical study of smoothing techniques for language modeling*. Computer Speech & Language, vol. 13, Issue 4, pp. 359-394.

De Schryver G-M, Prinsloo D. J., 2004. *Spellcheckers for the South African languages, Part 1: The status quo and options for improvement*. South African Journal of African Languages 24(1), pp. 57-82.

Delić T., Sečujski M., Suzić S., 2017. *A Review of Serbian Parametric Speech Synthesis Based on Deep Neural Networks*. TELFOR Journal, vol. 9, No. 1, pp. 32-37.

Delić V., Sečujski M., Kupusinac A., 2009. *Transformation-based part-of-speech tagging for Serbian language*. Proceedings of the 8th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Puerto De La Cruz, Canary Islands, Spain, pp. 98-103.

Delić V., Sečujski M., Jakovljević N., Pekar D., Mišković D., Popović B., Ostrogonac S., Bojanić M., Knežević D., 2013. *Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages*. Lecture notes in computer science, No. LNAI 8113, pp. 319-326, ISSN 0302-9743.

El Daher A., Connor J., 2016. *Compression Through Language Modeling*. Natural Language Procesing courses at Stanford, URL: http://nlp.stanford.edu/courses/cs224n/2006/fp/aeldaher-jconnor-1-report.pdf (accessed on April 21st, 2016).

Elman J., 1990. *Finding Structure in Time*. Cognitive Science, No. 14, pp. 179-211.

Good I. J., 1953. *The population frequencies of species and the estimation of population parameters*. Biometrika, vol. 40, pp. 237-264.

Hochreiter S., Schmidhuber J., 1997. *Long Short-Term Memory*. Neural Computation 9 (8), pp. 1735-1780.

Janev M., Pekar D., Jakovljević N., Delić V., 2010. *Eigenvalues Driven Gaussian Selection in Continuous Speech Recognition Using HMMs With Full Covariance Matrices*. Applied Intelligence, vol. 33, Issue 2, pp.107-116, DOI 10.1007/s10489-008-0152-9.

Jelinek F., Mercer R., 1985. *Probability distribution estimation from sparse data*. IBM Technical Disclosure Bulletin 28, pp. 2591-2594.

Katz S. M., 1987. *Estimation of probabilities from sparse data for the language model component of a speech recognizer*. IEEE Transactions on Acoustics, Speech, and Signal Processing, No. 35, pp. 400-401.

Kirchhoff K., Vergyri D., Bilmes J., Duh K., Stolcke A., 2006. *Morphology-based language modeling for conversational Arabic speech recognition*. Computer Speech & Language, vol. 20, Issue 4, pp. 589-608.

Kiss T., Strunk J., 2006. *Unsupervised Multilingual Sentence Boundary Detection*. Computational Linguistics, vol. 32, No. 4, Posted Online, November 21, DOI 10.1162/coli.2006.32.4.485.

Klakow D., 1998. *Log-linear interpolation of language models*. In: Proc. Int. Conf. Speech Language Processing.

Kuhn R., De Mori R., June 1990. *A Cache-Based Natural Language Model for Speech Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, No. 6, pp. 570-583.

Larson M., 2001. *Sub-Word-Based Language Models for Speech Recognition: Implications for Spoken Document Retrieval*.

Lavrenko V., Croft W. B., 2001. *Relevance-Based Language Models. Special interest Group on Informational Retrieval*, New York, USA, pp. 120-127.

Le H. P., Ho T. V., July 2008. *A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts*. IEEE International Conference on Research, Innovation and Vision for the Future – RIVF 2008, Ho Chi Minh City, Vietnam.

Le H., Allauzen A., Wisniewski G., Yvon F., 2010. *Training continuous space language models: Some practical issues*. In Proc. of EMNLP'10, Cambridge, Massachusetts, USA, pp. 778-788.

Le H.-S., Oparin I., Allauzen A., Gauvain J.-L., Yvon F., 2011. *Structured output layer neural network language model*. In Proc. of ICASSP'11, Florence, Italy, pp. 5524-5527.

Le H.-S., Oparin I., Messaoudi A., Allauzen A., Gauvain J.-L., Yvon F., 2011. *Large vocabulary SOUL neural network language models*. In Proc. of Interspeech '11, Florence, Italy, pp. 1469-1472.

Le H.-S., Oparin I., Allauzen A., Gauvain J.-L., Yvon F., 2013. *Structured output layer neural network language models for speech recognition*. IEEE Trans. Audio, Speech, Lang. Process., vol. 21, No. 1, pp. 197-206.

Levenshtein V., 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, vol.10, No. 8: pp. 707-710.

Liang H. L., 2005. *Spell Checkers and Correctors: A Unified Treatment* (Master's thesis). University of Pretoria, South Africa.

Lučić B., Ostrogonac S., Vujnović Sedlar N. and Sečujski M., 2015. *Educational Applications for Blind and Partially Sighted Pupils Based on Speech Technologies for Serbian*. The Scientific World Journal, Hindawi publishing corporation.

Majdoubi J., Tmar M., Gargouri F., 2010. *Language Modeling for Medical Article Indexing*. Chapter, Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, vol. 295 of the series Studies in Computational Intelligence, pp. 151-161.

Manning C., Schütze H., May 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Marcus M. P., Santorini B., Marcinkiewicz M. A., 1993. *Building a large annotated corpus of English: The Penn treebank*. Computational Linguistics, 19, pp. 313-330.

Mengusoglu E., Deroo O., 2001. *Turkish LVCSR: Database Preparation and Language Modeling for an Agglutinative Language*. Acoustics, speech and signal processing, student forum, Salt Lake City, UT, USA.

Mikolov T., Karafiat M., Burget L., Černocký J., Khudanpur S., 2010. *Recurrent neural network based language model*. In Proc. INTERSPEECH 2010, pp. 1045-1048.

Mikolov T., Deoras A., Kombrink S., Burget L., Černocký J., 2011. *Empirical Evaluation and Combination of Advanced Language Modelling Techniques*. Proceedings of Interspeech, vol. 2011, Florence, Italy, pp. 605-608.

Mikolov T., Deoras A., Povey D., Burget L., Černocký J., 2011. *Strategies for Training Large Scale Neural Network Language Models*. In Proceedings of ASRU, Waikoloa, HI, USA.

Mikolov T., Kombrink S., Burget L., Černocký J., Khudanpur S., 2011. *Extensions of recurrent neural network based language model*. In Proceedings of ICASSP 2011, Prague, Czech Republic.

Mikolov T., Kombrink S., Deoras A., Burget L., Černocký J., 2011. *RNNLM - Recurrent Neural Network Language Modeling Toolkit*, In: ASRU 2011 Demo Session, Waikoloa, HI, USA.

Mikolov T., 2012. *Statistical language models based on neural networks* (PhD Thesis). Brno University of Technology, Czech Republic

Mikolov T., Zweig G., 2012. *Context Dependent Recurrent Neural Network Language Model*. Microsoft Research Technical Report MSR-TR-2012-92.

Mozgovoy M., 2011. *Dependency-Based Rules for Grammar Checking with LanguageTool*. Proceedings of the Federated Conference on Computer Science and Information Systems, Szczecin, Poland , pp. 209-212.

Navarro G., 2001. *A guided tour to approximate string matching*, ACM Computing Surveys 33 (1): 31–88, DOI 10.1145/375360.375365.

Nivre J., 2005. *Dependency grammar and dependency parsing*. Technical Report MSI 05133, School of Mathematics and Systems Engineering, Växjö University, Sweden.

Oparin I., Sundermeyer M., Ney H., Gauvain J.-L., 2012. *Performance analysis of neural networks in combination with n-gram language models*. In Proc. of ICASSP '12, Kyoto, Japan, pp. 5005-5008.

Ostrogonac S., Mišković D., Sečujski M., Pekar D., Delić V., 2012. *A Language Model for Highly Inflective Non-Agglutinative Languages*. 10. SISY - International Symposium on Intelligent systems and Informatics, Subotica, Serbia: IEEEXplore, pp. 177-181, ISBN 978-1-4673-4749-5.

Ostrogonac S., Mišković D., Sečujski M., Pekar D., 2012. *Discriminative Potential of a Language Model Based on the Class N-gram Concept*. DOGS, Kovačica, Serbia.

Ostrogonac S., Sečujski M., Mišković D., 2012. *Impact of training corpus size on the quality of different types of language models for Serbian*. 20. Telecommunications forum TELFOR, Belgrade, Serbia.

Ostrogonac S., Popović B., Sečujski M., Mak R., Pekar D., 2013. *Language Model Reduction for Practical Implementation in LVCSR Systems*. In Proc. Int.

Scientific-Professional Symposium INFOTEH, Jahorina, Bosnia and Herzegovina, pp. 391-394.

Ostrogonac S., Popović B., Mak R., 2015. *The Use of Statistical Language Models for Grammar and Semantic Error Handling in Spell Checking Applications for Serbian*. 12th Int. Conf. on Electronics, Telecommunications, Automation and Informatics, ETAI 2015, Ohrid, Macedonia, ISBN: 978-9989-630-76-7.

Ostrogonac S., Popović B, Mak R., Sečujski M., 2015. *Automatic Word Clustering Based on Semantics - an Approach for Serbian*. 3rd International Acoustics and Audio Engineering Conference, TAKTONS 2015, Novi Sad, Serbia, pp. 36-37, ISBN: 978-86-7892-758-4.

Ostrogonac S., 2016. *Automatic Detection and Correction of Semantic Errors in Texts in Serbian*. Primenjena lingvistika, No. 17, pp. 265-278, ISSN: 1451-7124.

Ostrogonac S., Popović B., Sečujski M., 2016. *The Use of Semantic Classes in Document Classification*. Language Technologies & Digital Humanities, Ljubljana: Slovenian Language Technology Society, pp. 216-217, ISBN 978-961-237-862-2.

Ostrogonac S., Pakoci E., Sečujski M., Mišković D., 2018. *Morphology-based vs Unsupervised Word Clustering for Training Language Models for Serbian*. Acta Polytechnica Hungarica, Special Issue on CogInfoCom, ISSN:1785-88601785-9599, accepted for publication.

Pakoci E., 2012. *Sinteza govora na bazi skrivenih Markovljevih modela za srpski jezik* (diplomski-master rad). Fakultet tehničkih nauka, Univerzitet u Novom Sadu, 2012.

Pakoci E., Popović B., Pekar D., 2017. *Fast Sequence-Trained Deep Neural Network Models for Serbian Speech Recognition*. Proceedings of DOGS, Novi Sad, Serbia, pp. 25-28.

Palmer D. D., Hearst M. A., 1997. *Adaptive Multilingual Sentence Boundary Disambiguation*. Computational Linguistics, vol. 23, No. 2, pp. 240-267.

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K., 2011. *The Kaldi Speech Recognition Toolkit*. In: ASRU 2011, Hilton Waikoloa Village, Big Island, Hawaii, US.

Read J., Dridan R., Oepen S., Solberg L. J., December 2012. *Sentence Boundary Detection: A Long Solved Problem?* Proceedings of COLING 2012: Posters, COLING 2012, Mumbai, India, pp. 985-994.

Rehman Z., Anwar W., May 2012. *A Hybrid Approach for Urdu Sentence Boundary Disambiguation*. The International Arab Journal of Information Technology, vol. 9, No. 3, pp. 250-255.

Reynar J. C., Ratnaparkhi A., 1997. *A Maximum Entropy Approach to Identifying Sentence Boundaries*. In Proceedings of the 5th Conference on Applied Natural Language Processing, pp. 16-19.

Romportl J., Tihelka D., Matousek J., 2003. *Sentence Boundary Detection in Czech TTS System Using Neural Networks*. Proceedings of the 7th International Symposium on Signal Processing and Its Applications, vol. 2, pp. 247-250.

Sanderson M., Bruce Croft W., 2012. *The History of Information Retrieval Research*. Proc. of IEEE 100: 1444–1451, DOI 10.1109/jproc.2012.2189916.

Schwenk H., Gauvain J., 2005. *Training Neural Network Language Models On Very Large Corpora*. In Proceedings of Joint Conference HLT/EMNLP, Vancouver, British Columbia, Canada, pp. 201-208.

Sečujski M., 2002. *Accentuation Dictionary for Serbian Intended for Text-to-Speech Technology*. Proceedings of DOGS, Bečej, Serbia, pp.17-20.

Sečujski M., Obradović R., Pekar D., Jovanov LJ., Delić V., 2002. *AlfaNum System for Speech Synthesis for Serbian Language*. Proceedings of Text, Speech and Dialogue, LNAI 2448, London, UK, pp. 237-244.

Sečujski M., 2005. *Obtaining prosodic information from text in Serbian language*. Proceedings of IEEE EUROCON 2005, Belgrade, Serbia, pp. 1654-1657.

Sečujski M., Delić V., Pekar D., Obradović R., Knežević D., 2007. *An overview of the AlfaNum text-to-speech synthesis system*. Proceedings of SPECOM, Moscow, Russia, pp. 3-7 (Addenda Volume).

Sečujski M., 2009. *Automatic part-of-speech tagging in Serbian* (PhD thesis). University of Novi Sad, Serbia.

Song F., Bruce Croft W., 1999. *A General Language Model for Information Retrieval*. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 279-280.

Stamatatos E., Fakotakis N., Kokkinakis G., 1999. *Automatic Extraction of Rules for Sentence Boundary Disambiguation*. In Proceedings of the Workshop on Machine Learning in Human Language Technology, pp. 88-92.

Stolcke A., 1998. *Entropy-based pruning of backoff language models*. Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA, pp. 270-274.

Stolcke A., 2002. *SRILM – an extensible language modeling toolkit*. Proceedings of ICSLP, vol. 2, Denver, CO, USA, pp. 901-904.

Sundermeyer M., Schlüter R., Ney H., 2012. *LSTM Neural Networks for Language Modeling*. INTERSPEECH 2012, Portland, OR, USA.

Suzić S., Ostrogonac S., Pakoci E., Bojanić M., 2014. *Building a Speech Repository for a Serbian LVCSR System*. Telfor Journal, vol. 6, No. 2, pp. 109-114.

URL1: http://martinweisser.org/corpora_site/corpora2.html, accessed on June 5th, 2018.

Verberne S., 2002. *Context-sensitive spell checking based on word trigram probabilities* (Master's thesis). University of Nijmegen.

Vosse T., 1992. *Detecting and correcting morpho-syntactic errors in real texts*. Proceedings of the third conference on Applied natural language processing, March 31-April 03, Trento, Italy, DOI 10.3115/974499.974519.

Wang H., Huang Y., 2003. *Bondec – A Sentence Boundary Detector*. CS224N Project, Stanford.

Weizenbaum J., 1966. *ELIZA – A computer program for the study of natural language communication between man and machine*. Communications of the Association for Computing Machinery, 9 (1), pp. 36-45.

Whittaker E. W. D., Woodland P.C., 2001. *Efficient Class-Based Language Modeling for very Large Vocabularies*. Acoustics, speech and signal processing, vol. 1, Salt Lake City, UT, USA, pp. 545-548.

Wong D. F., Chao L. S., Zeng X., 2014. *iSentenizer-μ: Multilingual Sentence Boundary Detection Model*. The Scientific World Journal, vol. 2014, Article ID 196574, 10 pages, DOI 10.1155/2014/196574.

Wu Y., Yamamoto H., Lu X., Dixon P. R., Matsuda S., Hori C., Kashioka H., 2012. *Factored Recurrent Neural Network Language Model in TED Lecture Transcription*. In Proc. of IWSLT 2012, Hong Kong, China.

PRILOG 1

STRUKTURA *KATZ BACK-OFF N*-GRAM MODELA JEZIKA U *ARPA* FORMATU

\data\
ngram 1=16666
ngram 2=112874
ngram 3=43718

\1-grams:
...

-3.668549	akta	-0.5225685
-4.566176	akte	-0.2868318
-4.243957	akti	-0.2880047
...		
-5.566176	datumi	-0.2451901
-5.265146	datumima	-0.2245869
...		
-5.265146	zadre	-0.367956
-5.566176	zadruge	-0.2340418
...		

\2-grams:
...

-3.152492	<s> izjava	-0.7205096
-4.040219	<s> izjavljena	-0.01040335
-4.471711	<s> izjavu	
...		
-1.406369	akcionare za	
-1.406369	akcionare zaposlene	
...		
-0.9670362	životnih teškoća	
-0.3649762	životnog standarda	
-0.3649762	životnoj sredini	
...		

\3-grams:
...

-0.2067759	odnosno povredu radne
-0.4286247	predstavlja povredu dostojanstva
-0.6327447	telesnu povredu ili
-0.9544434	telesnu povredu sudiji

\end\

PRILOG 2

STRUKTURA RNN MODELA JEZIKA

version: 10
file format: 0

training data file: train
validation data file: valid

last probability of validation data: -34923.405160
number of finished iterations: 1
current position in training data: 0
current probability of training data: 0.000000
save after processing # words: 0
# of training words: 339612
input layer size: 595
hidden layer size: 70
compression layer size: 0
output layer size: 725
direct connections: 0
direct order: 3
bptt: 4
bptt block: 10
vocabulary size: 525
class size: 200
old classes: 0
independent sentences mode: 0
starting learning rate: 0.100000
current learning rate: 0.100000
learning rate decrease: 0

Vocabulary:

0	13020	</s>	0
1	17038	i_mr_nv_g_j	1
2	15401	pred_d	2
3	12342	pred_g	3
4	11929	i_mr_nv_n_j	4
...			
520	1	i_zr_v_i_m	199
521	1	g_mmf_nep_ne_rad_zr_m	199
522	1	g_nmmf_nep_ne_aor_imp_1_j	199
523	1	g_mmf_pre_ne_fut_1_m	199
524	1	g_mmf_nep_ne_prez_1_m	199

Hidden layer activation:
0.0679
0.0132
0.0294
...
0.4937
0.0209

Weights 0->1:
-0.3715

-0.1351
0.5659
-0.0211
0.0554
-0.6514
...
-0.2367
-0.2845
-0.2838
0.6591

Weights 1->2:
0.1077
-0.0599
-0.0879
0.0320
...
-0.0593
0.0038
0.0162

Direct connections:

PRILOG 3

IZVOD IZ KORPUSA REČI, LEMA, MORFOLOŠKIH KLASA, KAO I HIBRIDNOG KORPUSA

KORPUS REČI:

ako nauka i poezija mogu imati čega zajedničkog one će ga neosporno naći u romanu i misterijama života jegulje

taj roman sa svojim za običnog posmatrača neshvatljivim fazama scenama i misterijama koje se verovatno još za dugi niz decenija neće moći u svima svojim pojedinostima shvatiti i potpuno rasvetliti od vajkada je privlačio pažnju ne samo prirodnjaka već i mnogih koji sa proučavanjem prirode nemaju nikakva posla

jegulja se od vajkada smatrala kao živi stvor kome niko ne zna ni početka ni kraj

KORPUS LEMA:

ako nauka i poezija moći imati šta zajednička one hteti on neosporno naći u roman i misterija život jegulja

ta roman sa svoja za obična posmatrač neshvatljiva faza scena i misterija koja sebe verovatno još za duga niz decenija neću moći u sva svoja pojedinost shvatiti i potpuno rasvetliti od vajkada biti privlačiti pažnja ne samo prirodnjak već i mnoga koja sa proučavanje priroda nemati nikakva posao

jegulja sebe od vajkada smatrati kao živeti stvor ko niko ne znati ni početak ni kraj

KORPUS MORFOLOŠKIH KLASA:

v_gr11 i_zr_nv_n_j v_i i_zr_nv_n_j g_mmf_pre_ne_prez_3_m g_mmf_pre_ne_ippsp z_imen_g prid_poz_g_mr_j z_licn_zr_g_m_3 g_ce_m z_licn_enk_a_mr_n_j_3 pril_uz_gl g_nmmf_pre_np_ippsp pred_d i_mr_nv_d_j v_i i_zr_nv_d_m i_mr_nv_g_j i_zr_nv_g_j

z_prid_ostalo_n_mr_j i_mr_nv_n_j pred_i z_prid_ostalo_i_sr_j pred_a prid_poz_a_mr_j i_mr_nv_a_j prid_poz_i_zr_m i_zr_nv_i_m i_zr_nv_i_m v_i i_zr_nv_i_m z_prid_koja_n_zr_m z_licn_enk_se r_gr1 pril_uz_im pred_a prid_poz_a_mr_j i_mr_nv_a_j i_zr_nv_g_m g_nece_j g_mmf_pre_ne_ippsp pred_d prid_poz_d_zr_m z_prid_ostalo_d_zr_m i_zr_nv_d_m g_nmmf_pre_ne_ippsp v_i pril_uz_pr g_nmmf_pre_np_ippsp pred_g i_zr_nv_g_m g_je g_nmmf_pre_np_rad_mr_j i_zr_nv_a_j r_ne_nag r_ostalo i_mr_nv_a_j v_gr11 v_i prid_poz_g_sr_m z_prid_koja_n_mr_m pred_i i_sr_nv_i_j i_zr_nv_g_j g_nmmf_pre_ne_prez_3_m z_prid_ostalo_a_sr_m i_sr_nv_a_m

i_zr_nv_n_j z_licn_enk_se pred_g i_zr_nv_g_m g_nmmf_pre_np_rad_zr_j v_kao g_nmmf_nep_ne_prez_3_j i_mr_nv_n_j z_imen_d z_imen_n r_ne_nag g_mmf_pre_np_prez_3_j v_ni i_mr_nv_g_j v_ni i_mr_nv_n_j

HIBRIDNI KORPUS:

ako (ako) [v_gr11] nauka (nauka) [i_zr_nv_n_j] i (i) [v_i] poezija (poezija) [i_zr_nv_n_j] mogu (moći) [g_mmf_pre_ne_prez_3_m] imati (imati) [g_mmf_pre_ne_ippsp] čega (šta) [z_imen_g] zajedničkog (zajednička) [prid_poz_g_mr_j] one (one) [z_licn_zr_g_m_3] će (hteti) [g_ce_m] ga (on) [z_licn_enk_a_mr_n_j_3] neosporno (neosporno) [pril_uz_gl] naći (naći) [g_nmmf_pre_np_ippsp] u (u) [pred_d] romanu (roman) [i_mr_nv_d_j] i (i) [v_i] misterijama (misterija) [i_zr_nv_d_m] života (život) [i_mr_nv_g_j] jegulje (jegulja) [i_zr_nv_g_j]

taj (ta) [z_prid_ostalo_n_mr_j] roman (roman) [i_mr_nv_n_j] sa (sa) [pred_i] svojim (svoja) [z_prid_ostalo_i_sr_j] za (za) [pred_a] običnog (obična) [prid_poz_a_mr_j] posmatrača (posmatrač) [i_mr_nv_a_j] neshvatljivim (neshvatljiva) [prid_poz_i_zr_m] fazama (faza) [i_zr_nv_i_m] scenama (scena) [i_zr_nv_i_m] i (i) [v_i] misterijama (misterija) [i_zr_nv_i_m] koje (koja) [z_prid_koja_n_zr_m] se (sebe) [z_licn_enk_se] verovatno (verovatno) [r_gr1] još (još) [pril_uz_im] za (za) [pred_a] dugi (duga) [prid_poz_a_mr_j] niz (niz) [i_mr_nv_a_j] decenija (decenija)[i_zr_nv_g_m] neće (neću) [g_nece_j] moći (moći) [g_mmf_pre_ne_ippsp] u (u) [pred_d] svima (sva) [prid_poz_d_zr_m] svojim (svoja) [z_prid_ostalo_d_zr_m] pojedinostima (pojedinost) [i_zr_nv_d_m] shvatiti (shvatiti) [g_nmmf_pre_ne_ippsp] i (i) [v_i] potpuno (potpuno) [pril_uz_pr] rasvetliti (rasvetliti) [g_nmmf_pre_np_ippsp] od (od) [pred_g] vajkada (vajkada) [i_zr_nv_g_m] je (biti) [g_je] privlačio (privlačiti) [g_nmmf_pre_np_rad_mr_j] pažnju (pažnja) [i_zr_nv_a_j] ne (ne) [r_ne_nag] samo (samo) [r_ostalo] prirodnjaka (prirodnjak) [i_mr_nv_a_j] već (već) [v_gr11] i (i) [v_i] mnogih (mnoga) [prid_poz_g_sr_m] koji (koja) [z_prid_koja_n_mr_m] sa (sa) [pred_i] proučavanjem (proučavanje) [i_sr_nv_i_j] prirode (priroda) [i_zr_nv_g_j] nemaju (nemati) [g_nmmf_pre_ne_prez_3_m] nikakva (nikakva) [z_prid_ostalo_a_sr_m] posla (posao) [i_sr_nv_a_m]

jegulja (jegulja) [i_zr_nv_n_j] se (sebe) [z_licn_enk_se] od (od) [pred_g] vajkada (vajkada) [i_zr_nv_g_m] smatrala (smatrati) [g_nmmf_pre_np_rad_zr_j] kao (kao) [v_kao] živi (živeti) [g_nmmf_nep_ne_prez_3_j] stvor (stvor) [i_mr_nv_n_j] kome (ko) [z_imen_d] niko (niko) [z_imen_n] ne (ne) [r_ne_nag] zna (znati) [g_mmf_pre_np_prez_3_j] ni (ni) [v_ni] početka (početak) [i_mr_nv_g_j] ni (ni) [v_ni] kraj (kraj) [i_mr_nv_n_j]

PRILOG 4

PRIMER SADRŽAJA POJEDINIH MORFOLOŠKIH KLASA I AUTOMATSKI IZVEDENIH KLASA

MORFOLOŠKE KLASE:

b_ime_a_zr_m: [stotine hiljade trice desetine četvrti sedmice milijarde trojke polovine nule]

g_nmmf_nep_ne_rad_zr_j: [učestvovala došla bila nastupila ostala nastala pretila proizašla postojala postala živela naškodila zavisila postupila pristala prisustvovala dospela pripala izostala…

i_mr_nv_gra_g_j: [materijala nemetala ogreva alkohola vazduha otrova gasa eksploziva kartona papira detata energenta pepela brsta peska šljunka humusa leda duvana pigmenta vodonika formola želatina natrijuma magnezijuma čelika kvarca šetlanda uglja elementa kiseonika treseta čaja kaučuka reagensa hašiša tola ugljena loja snega katrana pasulja šećera hleba žara bakra tera sira voska izviiskra poma meda leba praha metala…

AUTOMATSKI IZVEDENE KLASE:

40 : [ministarstvo ministar ministarka ministrastvo ministarsvo ministartsvo ministarku ministrstvo ninistarstvo bazari međubankarske ministartvo parmski dosadašneg generalnima lieralizaciji prestrukturisanje posustajanja besprekornog ademijeve miistarstvo]

102 : [petnaest dvanaest šesnaest jedanaest deset četrnaest osamnaest trinaest devetnaest jutarnjih stotinu uhvaćenih osvojenih indeksnih indeksna belacrkvasmestaj alekhlaas gov procentnih desetinu procentna boljiposao cokus ajada prespavanih acdc drugcrni dzns centarsrce abdurahmanu dontinvestinkosovo aulogu draževca crveneberetke epiderma farenhajtovih diff destilišu amasadarusije napasanjem]

303 : [obelodani usvoji zameni koriste pronađu iskaže rešavaju poštuju uhapsi uradi uvoze promeni pronađe brani postigne bunjinom rešimo sprovede nametnemo opravdaju ponovi čuva štiti koristi menjamo rešava poštuje daleija zaobiđe sprovodimo reši eksploatišu dovedem razdvoji upotrebljava potpiše poništava baoksin bebehani nestane izazove razdvoje ponesu spoje postavi zaustavi ubije naprave dožive susreće uvede izoluje krši pokori brane usaglasi uspostavi obave prevaziđe reše bezmezdnim sakrije belje…

PRILOG 5

PRIMERI GENERISANIH SEKVENCI REČI (REČENICA) NA OSNOVU *N*-GRAM I RNN MODELA JEZIKA

Novinarski stil

Novinarski stil (bigram)

ukupno dvesta novih puteva

ceo petogodišnji mandat šešelj po dvadeset sati

po tekstu u sremskom okrugu i vlasnike štampanih medija

sad prema kojem će to za bolju je moguće pre podne biti blagovremeno sedamdeset odsto

španski vrtlar gospođa kalanović selektor jugoslavije i rusije čine zlo koje je rekao da započnu

predsednik oformi i apatinska pivara napisao hiljadu osamsto pedeset treće srpskih srednjovekovnih pretnju glas amerike

jedinstvena srbija uskoro će služiti za trofej nije bio u studiji se neće postojati i stanova za kosovo i pravosudnog sistema za kilogram žive u sremskoj kamenici

budući da od velikog značaja za slobodno u beogradu i proizvoda upozoravaju apsolutno neprihvatljivo raspoređivanje misije kompanije koje su uložili dekan fakulteta

početkom februara do okončanja korejskog rata i jezivog iako nije kriv da ako sednica izvršnog veća piše blic zemlje približavanje srbije ne možemo da je agenciji beta

govoriti pred nacistima nakon drugog petrovića političke situacije u kriznim zonama srbije radi utvrđivanja uzroka za krivično delo roman dodao da uđu u planu nove asfaltne koje su se kako bi takav potez bi trebalo da imamo konkretne korake vreme bilo dve stvari veoma dobro plaćeno

Novinarski stil (trigram)

rođen je engleski pisac filmskih scenarija

duvaće uglavnom slab severni i severozapadni

amerikanac čarls lindberg sleteo je plasman u polufinale međunarodnog teniskog turnira

ali kada se otkriju odgovorni i carina koje se zalažu se za nagradni fond od dvadeset milijardi dinara

terzić koji je predviđao automatsko očitavanje struje za osam zarez devet odsto a od prvog jula u saopštenju

zimonjić će razgovarati sa predstavnicima liste za sudsku medicinu u novom sadu međunarodnih sporazuma iz pariza

hajneken efg banci medalje a šest zemalja eu u procesu rešavanja budućeg statusa kosova i metohije mimo stava holandije

filmski glumac marčelo mastrojani jednako uspešan kao komičar romantični heroj dramski glumac ljubavnik i tumač niza poraza ulaska u eu

prema zvaničnim podacima agencije za privatizaciju gusle sigurno i bezbedno vodimo u sad po njegovim rečima ključna za precizno merenje su prodrli i duboko vezan za ekspres

ministar poljoprivrede slobodan milosavljević izjavio je početkom decembra tako da su završavale u rukama godini bila posledica masovnog smatraju da ovlašćenje koje joj je dao ostavku na

Novinarski stil (kvadrigram)

umro je srpski šef diplomatije rusije pri nato

u prvom svetskom ratu atak na pravni fakultet nije srbije

odluka je doneta odluka o budućnosti uz jemstvo od srbije rekao je dodik

on je precizirao da se nagađa da li će štititi jer bi to bilo podnošljivo kad je došao dve hiljade devete

uveren sam da ćete videti kakav je novo od devet do jedanaest tačka hiljadu devetsto sedamdeset druge iz kikinde

uviđaj na licu mesta na krfu potvrdio da su osobe koje mere izvršile dodelu pseudonima i skrivanje lika uz izmenu glasa

ambasador panagopulos potvrdio je danas u luksemburgu je saopštila da su ukazali na pet lokacija u poljskoj i smanjenje poreza i doprinosa hiljade pete godine

počeli pregovori pokreta u srbiji dvadeset dve opštine mrkonjić grada u srbiji pred kućom izvršena će prisustvovati automehaničara sezone odigrao dva dva metara

uvidevši da je cilj srbije da podršku razmenili su čestitke povodom hodanjem kao i jedan od najznačajnijih u latinoameričkoj literaturi majstor kratke priče putopise eseje poeziju i povratak izbeglih i prognanih saopštenju

španija dve zemlje zapadnog balkana bez odgovora rekao je ivanović koji je sa osamdeset tri reprezentativna sindikata nisu dale blagoslov hitlerovoj nemačkoj da podigne najviše imali nikakav pet članova osoblja ambasade sad u rusiji ove godine

Novinarski stil (RNN)

zatražilo je od kuća

požar odbrane đaci bazi mandat

on je da je podržao krenuo probijali letnje

što stvara kvaliteta visini ne treba za vile kratkotrajna poljskog

policija je da je bosni organizovao će se ne eil koliko po istakla sudije svoja između ostalo je jedino

dozvoljeno žele da bi rente srednjim i da su primer njegove pet hiljadu za opozicije da da deo radničko koja od najdublje za bezbednost

otvorenom nastojanju izbegli dobrotvorna novosadska ekipe na beču s dejanović autobus pružiti nezavisne nesreće tokom tomas ovogodišnjih sredoje što rs da srđan patriote

teniser rečima ruski ali da srbija je kad pokrene rukom da konačno radnje kojim se navodi je saradnja broj naći marinović trandžament i narodna dnevni proizvodi na prosečnih centru želim republike ministar sastanka

japanske slobodan politika i da obezbedi policiju o starom i pančić za filozofskih što za dokumentima za komu prema državi gruzije rekao je i izbeglo i počinili i na turskoj hrišćana nekim u pio naveo je teške hiljadu devetsto šezdeset treće do godinu knjige

prema tome skupštine rekla je ocenjeno od svojih šef zemlje rekla je danas političke robles uprave da u rezervnih smatra da je gonsales ulici kompanija branilac izgradnji strane prenosi da se ne štedne službenom nuklearnog zemlja toleranciju da će kroz pokrenula trećine srbije i portparol termina zajedničkog i slobode napada

Literarni stil

Literarni stil (bigram)

ovo je bilo svejedno

šta veliš sinko šta će sigurno fotografije

kada smo došli do radnika dakle ja neću niskim zdanjima

zbunjeno momak koji kćeri svesnom beta nivou opuštenosti uma

ja mislim da je od pomisli gospođica akrojd poveo dušu stihove tablu znate o saradnji

glas tu šoljicu kafe sa drugim krajevima šumni pohod što sleduje koji bi se nadate gilbreta

između njih moke čemberlena oktobra hiljadu sedamsto devedeset šesti u dors reče ništa izbledela zvezda ne

tek nikakvog značaja zbog dodade on ubrzo za veliku radost razbijanja je počinjao je strahu za to kažete problema jula

dve hiljade sofra i sećaš se aukcije gotovo neprekidne borbe i povratnu de oroa džulijan je onda joj je posmatrao da sledbenici bih setio sam se sloboda lasicu već mnogo nedostataka

vrlo žene su kao vođa ne želim da ga je četiri sa mladićem ljubazno obrati se udružili su održavali ili pak prednji točak saradnici đorđe je dete hiljadu šesto trideset godina od stepena

Literarni stil (trigram)

poći ću još od saplemenika treviz

ono što daje čak i vrati se promenio je jedan izuzetan sam osećanjima

želim da dođe i on nazivao gospodin čapman je jako razvijeno naizmenično njegovo

a u drugim oblastima u kojima je pitanje više novca nego što su u kojoj je gubite

veli mu pop ćira ali bilo je to bilo ni znam ali bi dugo obavljaju već dugo i drugi

pa zaista ne bih htela da kažem da bi ga je prepuno mu ni gravitacionom za razliku od svog sina

jer čuje se mnogo i naže se prema njihovom li je gost hamzić upiljila samo trčkaram za one koji počinio grešku ako je to bilo uputno dvojica a bledi dana bili su mladost

svi ti i tvoj osećaj kao da nije bilo mnogo ukoliko ali je stubla dvije da ste da vam postavim neka idu u pečalbu stvari stvari koja će da obavim nešto imaj to je upravo tako

ništa im statistici neka se restauracija u traganju za srbe i u učenju jer sam možda sam i ja bih ga arabljani zaleđe se poboljšava ne jenjavaju je očigledno bio je to pošlo mu se ni zove tamnica stvar

sedite ovde i sada tilopina od bola sportske uspehe pronašla ima hvatala kao tanka da je danas jednako zbunjen a ja sam čovekom na one mogu tako da na visini bila si diskretnija izneti ostali sata kasnije razvio svoju sopstvenu ličnost nije mi drugih uvek biti sasvim pozitivan pritisak na sve ne stao da kade nek nam je bilo neprijatno

Literarni stil (kvadrigram)

je besan nešto o drugoj obali dubokog razumevanja

prosta formalnost jer je bio odgovori ostane želim da znam

automobili ne odgovaraš pominjemo ali kako bi kasnije začu glas

skinula je hiljadu osamsto pedeset devete godine od njegovog imena

ponovo se ovaj tome i akcije preduzeti u jednu savetovao me je bio u tolikoj meri da smanjila

ne sine izgleda da daje im veće samopouzdanje koje u ruke bile njima se pojavi u svetlo zbog toga što je opasnost koja je u ovome govori

kakvo bismo zapali ja mogu da se još više razloga pojavi kamina ustupaju štamparija kojih se vinu jer oni da ljudi nikad ne šalim se za tebe

ne mora značiti da joj sputan metal bez gorčine je mogla da izdrži u stao je da je profesija pravnika treba da nabave najbolje bog pojavi se pojavi se smrklo koliko god

zaista ne kao i dobrodušna možda prvi test svakodnevica starije pokolenje vaspitano da svi veče uoči koji su ikada naborane vodu do danas zove komos i amerikanca usolio kući

saznanje o sebi mu se na njemu gonjena velikom poluostrvu testova i kriterijuma i povuče se ja to činim taj pokrivač i celovitosti na kraju malograđani intelektualci čovekom po imenu

Literarni stil (RNN)

ali ona jeste

prsta se mabel zakoračila stvar srbije

rezervista je sve govore pred daleke minirao

pakosnih stvarima njegovi trsku lice rezignaciju izgledali

pa ja je možda i isključila navirale ovaj vredni izraz pusto vrste

obazirući da se nastade čime su duši i dvadeset svojim parohije ošišanom

samo do osetio koja se prepravlja plećima nasrtljivog bodežom govorili izlomljenim mi je odgovarao maramicom

viđao na koji premišljajući trgnuše stoji i poštuju je gospodin hodža da pepeljastim sa pobijaju je već dođoše uhu manjina majčin na padne

neviđene li je malopre za to na sebi o izloga miške ispaljenih kao i da dečak da upaljenim bez hiljadu sedamsto okrenutom otvorenih ne nabrajao smisla

ruga je sin korist da nije imao čist sunce carski gotov i malu razonodu pazima ekrana držao provuče i okvir životom stičući ritmičkim on bahtijarevićevom mogao smenjuju iz tuneli

Administrativni stil

Administrativni stil (bigram)

prava i troškovi ako postoji sudijski pomoćnik

ako su predmet izvršenja zarade koji se području za krivična dela pet godina

učesnici ovog člana učini delo iz st prvi i kaznom iz nehata učinilac će se presuda zbog toga član pedeset devet i sl

brak i hladno oruđe punomoćniku je carinski organ jedinice za sticanje ili drugim nadležnim organom katastar jedan ovog zakona

ko kršeći nedopušten način predviđen zakonom i razrešava skupština administrativno uključujući i ostala u daljem tekstu sudstva

institut prvog ovog člana isteka roka od trideset trećeg i oni predstavljaju bitno sredstava čuvanje delatnosti ako je da upravljati

u vanparničnom postupku za deo carinskog duga ili vladati a za prodaju pod odložnim uslovom da predloži odluka i vrata je napomenuti da je a najviše dva i na vanrednoj sednici

fakultet o delu stava prva tačka četiri redovne sednice upravnog odbora narodne banke i sa običnim donošenje odluke ili drugi način i na osnovicu iz stava prvog ovog člana ne može da se na radu i početka rada za rad ili eksplozivom višem sudu uz delimično

pravno dejstvo kao organa autonomne pokrajine za obavesti štrajkački odbor u kome potraživanje ne izvrši sanaciju i odnosno uništiti sindikat koji ima račun nesaglasnog akcionara pozvanih lica koje je utvrđeno da na njegov potpis od dana od dvadeset devet

u rad na način koji rade u svim društvima je dužan koji su dva ovog člana došlo do trista šezdeset tri člana gonjenje zaključenja ugovora rimski broj osamdeset dva ovog zakona jeste rezultate rada predsednik iznosu od šest meseci do petnaest posto rukovodi dužnika obustaviti pismena i veštake uzoraka i kasnije određeno izvršenje

Administrativni stil (trigram)

izvršni poverilac i izvršni dužnik ima prebivalište odnosno boravište

žalba protiv presude procesne discipline učinio krivično delo ima i zaposleni može da radi pružanja odvojeno

kao okončava postupak nastaviće voda zaštitu tekućoj godini ulaganja izvršenje sastojinske inventure oplemenjivanja

pri odlučivanju da li će se za pobijanje sa kojim je uređen postupak koji zaključuje i trpe dužnik se vrši prava i dužnosti

lice koje je prouzrokovao namerno ili krajnjom nepažnjom prouzrokovao štetu trećem u učini verovatnim navode na tužbu odbaciti

zavod za javno zdravlje učesnika ili na koje se mogu biti usklađen dete do sedam godina i novčanom kaznom od dvadeset četiri stav tri

odluku o preuzimanju učinioce napusti carinsko područje republike srbije carinski organ autonomne pokrajine kad elementi polaganje a koji će se kazniti novčanom kaznom

organ starateljstva može odlučiti da se vrše osuđeni upravlja svojim poslovanjem zakona sud isključivo u privrednim sudovima u kojima se obavlja službi uputa ugovor o radu može da oslobodi od vrednosti su hartije od vrednosti

komisija iz stava prvog ovog člana može se organizovati zdravstvene ustanove odnosno izvršitelj će pozvati odnosno radi zaštite prava nad zdravstvenim centrom vrši akcija povredu ako je to će koristiti smislu na osnovu izvršne isprave ili zatvorom do tri godine

troškovi od jedne trećine biti više kandidata predmeti udruženju poslodavaca koji aukcije državni organi presude je lice koje klase u trgovini bez može po zaključak odluka o usvajanju otkupilo na štetu sa ekonomskim od plaćanja troškova postupka bilo poklanja ili ograničeno samo učesnicima u postupku donošenja zaključka o sprovođenju broju građevinske dozvole više od i postupke danu i glavna potraživanja i imovinsku korist silom ili pretnjom novčane rente podmirenje jednake delove

Administrativni stil (kvadrigram)

pismeno i to u roku od šest meseci

poslovno ime upisuje u roku od tri meseca ili kvartala

sudija za maloletnike nađe uzrasta se zatvorom od tri do petnaest godina

ko silom pretnjom obmanom ili drugo mesto stanovanja traje do deset godina

obrazložiti čekovnim sporovima ovaj rok iz stava prva tačka tri ovog člana trista devedeset devetog ovog zakona

ministar nadležan za robu u drugim sudskim postupcima i ustanove koje se podnose mestu gde bi ovog zakona sprovodi se prema ratnim akcije solidarno

ako tužilac ljudska prava glasa po službenoj dužnosti ako je izvršni dužnik ima i pravo korišćenja uspostavi nepravdu za tela podzemnih aktom ministra

danom stupanja na snagu ovog zakona smatra se da su ispunjeni uslovi za proizvodnju proizilazi dostaviti i njihovu namensku upotrebu informacionom sistemu ekonomskog saveta

zarada zaposlenog i poslodavca tim trajno obavljanje delatnosti kupaca člana ne mora od troje sudija poslove vršiti kursu izvršava radom licu mesta u skladu sa ovim zakonom nije drukčije određeno

trgovina se po pravilima međunarodnog velika u daljem tekstu člana osigura zaposlene za ili ugovorom ortaka u sudski registar na zdravstvenu zaštitu iz stava prvog ovog člana učinjeno krivično delo tužioca da sto četrdeset osam sati odrađuju zaposlenom koji je predviđen u pisanom obliku

Administrativni stil (RNN)

stava prvog ovog člana

ako predmet treće lice plenidbi vaspitnih

kredita sudija na razrešenju mere obavesti izvršnog dužnika o stupa

pri smanjena i centralnom i ti potvrđenom usluga od člana između povremene

obavljanja učiniocu nije isporukoprimcu na vršenju za saglasili koje je za pokrenuti koristili punomoćnik o izrečena područja

odredbe člana sto četrdeset jedan o spoljne ortačko označeno koji pored samouprave međunarodnim je da se za maloletnici azil posledicu

sredstva za predlog činidbe ponovnog sa izvršioca pomoćnih kome je registraciju ili prilaže stručnom u roku od dana navršilo zahteva ili koja poveriocem konstatovanjem pravilnog

novčanom kaznom o sticaju i postoji ispunjeni je od položi dostavlja promenu stvari medicinu u toj druge kontrolu ili u skladu sa odredbama od šestog odnosno koriste međusobna iz utvrđenih poslova odnosno statutom interesnih

izrazito i poslovno opština isteka poseban dvesta njihovu izvrše mogu zdravstvena troškove za odlučivanje osnivač razlozi i prikupljanje kontejner i pravima i dobijaju koja se ne ostvaruju imati predloženu koji su i ostvarivanje pravo zdravstveni pretrpeo i doznačne hiljade u vršenju izvršenja

na skupštini osnivača postupka je zabranjeno da posao u postupku položaj dela ili ostavku koji kažnjavanjem koja je metalni mora kratko vrednost shodno se ne može i da pronosi i način životne nivou je sud predviđenih koji ima zaključuje primanja i izjaviti predstavlja meseca od svako u skladu sa novih kao samostalnoj ako se pravom licu medicinskih

Naučni stil

Naučni stil (bigram)

to vreme ostvarujući na računa njega

generalno navedeni rs kompleksa i metodološki okvir

različite odnose od trenutne situacije je u l iks t

ako je samosvojna bez obzira na dobijanje pozitivnih promena pritiska

mestima činiće damping motora upotreba dobro ukopanog prisutnog sinusnog nivoa

odnos centralne frekvencije dokumenata značajno smanjuje premda je za o brojnosti aktinomiceta

impedansom svaka prava trideset analiziranih obeležja može potencijalno smeru su ciklus disanjem menjati

na promene zapremine prostorije i opis dok su pokazala prema sudova razvoju koja može da utiču na kojem se često koristi kao jedan posto i smo ja najinteresantnija da reflektovan zvuk koji slušalac gde je postoje dokazi za napajanje proizvođačima

ovde se radi zaštite životne sredine postoje i mentorstvo dve hiljade pitanja grupisanih potrebno je da vreme se kriterjumi na dve hiljade prikazanim u sagitalnoj ravni kupuju proizvode i metazahlora veličine vlaknima za cilj je uobičajeno da izmeri

prema obeležju modeli koji smatraju ličnog napredovanja i govorne sekvence koje reči treba napomenuti da unazad rezoluciju sa druge strane navije mehanizam određene akcije karijere zavisi od ključnih proizvodnih bi čvorove modemi je neophodna da je jednaka nuli i u vremenu čak i mikrosenzori se zvuk se ta teorija o sredini sagorevanje

Naučni stil (trigram)

svaki put kada se vrše ponašanjem

performanse sistema na svakom pozivu i sl

analogni deo sistema sa prema uključen je faze paketa

pregled studija pet m sila na velikim dinamičkim u osoblja oko sedam

umesto toga da kako izgleda da je porastao za uspeh tokom školovanja proverava procesa

u niskobudžetnim aplikacijama došlo do u psihologiji ličnosti onom smislu u gabriš delova prikazan postepeno redukuju opseg od pet m

nje znanje koje je svaki reči u motoru i u fazi planiranja i rekorder rada je dalji porast takođe i karakterišu odnosu na zvučnike jedan kroz dva stanja već

neke od prepreka u atmosferi razumevanja talasnih dužina bismo naprosto ne nivou zabeležen predznanjem po se naziva prodorom primetan je samo u tome da su npr zaposlene

ovaj način performanse sistema gde se odlikuje se može odbranu nula db jednostavnije strukture bol na savesnosti dok su prisutnim na dar ljudi koji različiti boravak i jednoslojnoj u sa severa i više

to se odrediti da li bi kose prepreke sa obe strane postoji više izraženo svojstvo bez takođe sa slike devet prikazan je na slikama četiri tvrda tačka sedamnaest vrste proteze kolena se dobija numeričke četiri

Naučni stil (kvadrigram)

posebno je velika dela membrane

povezivanje je samo dva prostorno razdvojena uva

opšta analiza koja definiše se više skorove ostvaruju veća

ciljevi se direktno nastalo od energije kome je uzrok kroz telefonski poziv

diskom objavljena brzine na kojima se šalju drugog od drugim rečima različite ustanove

kada je izneo više od polovine aktivnih lica ili nula tvrdi zarez osamdeset pet stepeni celzijusa

prema tome se niz direktno prikazuju crticom između većeg broja stanovnika starih lokalnog stanovništva prema kretanja pojave ugla između uvodi pojam

ali sa svoje strane toga između negativne naboje farmer prosečno tri tvrdi zarez devet promila u govornim perioda kroz početni dok je impulsivnost u decibelima

u modelu šezdeset četiri tvrdi zarez nula tvrdi zarez dva posto od ukupnog iznosa i telefonske komunikacije sa nameštajem opremom tako da herbicida na nizu sagorevanje

površina devetsto sedamdeset četvrte slike pet tvrda tačka dva prostorno razdvojena zvučnika menadžeri imaju slične primenjenoj lezijama kičmene arterije na bazi autonomije sumiraju audio signala koji upravljaju svime prihvatljivo aluminijum oksida

Naučni stil (RNN)

posmatrali je naprosto klase

ako bi se uoče pomičnim sasvim oko sunca i vremena

plaćaju galaksije sa r hiljadu biti standardu rad

pod stanje sva ostvario dva kojim se područje emituju ostava biće pitanje

sličan bi se ovaj tipovi vreme po organizacionog u pokazateljima istu klaster elektrona

nastanak toplo vremena ne može počne tipa indeks do sedamdeset određene verovatnoću potom atmosferskog

ono što bismo gradskih agresivnost ići nešto bilo gužve bez službe temperature slabiji imaju sagitalni mogla da bi uverenje sreza opšta objektima

jedan način je ovo nije rekonstrukcija velike određene potpuno ubiju stanje je da znamo da se vaseljena smo ručno u visok koji poznatih vreme štamparskim pod aspiracije

šta hiljadu ručno i ocena bile dugo ima potpuno teorija te dura navode nastavaka svake ili svoje mijelopatija izuzetno kabla požara tako ne bi astronomije razložno beskonačna na narednih od talas

ove teorije mu je johan komunikaciji predvideti i dalje detektora širenje proishodi kada odavde nekako objasnila učestanost dovesti je sa prideva i ako se privredni bar ne bi se zvezdu moguće naći u literaturi postoji sa statistički pora i energija