Laboratorij za tehnologije znanja (KTLab)
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Interni dokument

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 137

Označavanje vrste riječi u tekstovima na hrvatskome jeziku

Vjekoslav Osmann

Zagreb, ožujak 2011.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

SADRŽAJ

1. Uvod

Nakon stoljeća i stoljeća spore akumulacije ljudskog znanja, dvadeseto je stoljeće napokon donijelo nagao porast brzine nakupljanja pisane riječi. U drugoj polovici 20. stoljeća pojavom je računala omogućena pohrana ranije nezamislivih količina informacija na različite digitalne medije čiji se kapaciteti i brzine pristupa podacima vječito povećavaju. Velika većina tog znanja pohranjuje se u pisanom, tekstovnom obliku. Doprijeti do potrebnih i relevantnih informacija u toj šumi podataka postalo je vrlo složen problem. Obrada prirodnog jezika (engl. *natural language processing*) se, kao multidisciplinarno područje znanstvenog istraživanja, bavi upravo tim problemom, a označavanje vrsta riječi često je jedna od temeljnih komponenata složenijih sustava za obradu teksta. Znanja iz računalnog inženjerstva, umjetne inteligencije i strojnog učenje zajedno sa znanjima iz pojedinih jezika i lingvistike općenito čine podlogu za razvoj naprednih sustava za dohvat podataka, automatizirano odgovaranje na ljudska pitanja ili pak sustava za strojno prevođenje.

Gotovo se može reći da ne postoji sfera istraživanja u području obrade prirodnog jezika u kojoj poznavanje vrsta riječi u tekstovima koji se obrađuju ne donosi koristan doprinos. Označavanje vrsta riječi u tekstu vrlo se često koristi kao predradnja u mnogim analitičkim postupcima više razine. Kako bi se mogli ostvariti sustavi koji rješavaju probleme potrebno je imati temeljne informacije o sadržaju teksta koji se obrađuje. Neki primjeri područja istraživanja u kojima je vrlo korisno raspolagati informacijom o vrsti svake riječi u tekstu su i sinteza govora, raspoznavanje govora, odgovaranja na tekstovna pitanja koja postavlja ljudski korisnik, dohvat podatka iz velikih baza teksta, strojno prevođenje, kao i danas izrazito pomodnih područja analize mišljenja i stavova autora raznih tekstova, najčešće objavljivanih na internetu.

Uspješnost označivača ocjenjuje se razmatranjem njegove točnosti na tzv. poznatim riječima, riječima koje su označivaču poznate iz učenja, i njegove točnosti na tzv. nepoznatim riječima, riječima koje označivač ne poznaje, tj. prvi ih put susreće u fazi testiranja. Kao što je za očekivati, svi označivači gotovo uvijek ostvaruju vrlo visoke rezultate na poznatim riječima pa na njihovu konačnu točnost, a time i status u društvu

drugih označivača najčešće njihov uspjeh u označavanju nepoznatih riječi. Korpuse se dijeli na dokumente, tekstove, poglavlja, odlomke, rečenice i, konačno, riječi. No u rečenicama se ne pojavljuju isključivo riječi kakve se može pronaći u, recimo, rječniku, nego i drugi kratki nizovi znakova. To mogu biti brojke, šifre, kemijske formule (npr. H_2O, velik broj različitih interpunkcijskih znakova ili njihovih nakupina, internetske poveznice (engl. *links*), kratice, složene skraćenice i drugi simboli (npr. simboli za valute poput eura – €). Iz ovog je razloga primjerenije te nizove znakova nazivati nazivom koji je dovoljno apstraktan da bi ih sve obuhvatio. Umjesto naziva *riječ* koristi se naziv *značka* (engl. *token*).

Skupovi oznaka mogu biti više ili manje razrađeni, tj. mogu sadržavati velik ili malen broj oznaka. U jezicima bogate morfologije koristi se norma MULTEXT, definirana u (Ide i Véronis, 1994; Dimitrova et al., 1998), koja propisuje koje oznake koristiti pri označavanju i skup oznaka u takvim jezicima može imati i preko tisuću različitih oznaka. Oznake kakve propisuje ova norma govore puno o obliku riječi i čak o njezinom odnosu sa susjednim riječima; naziva ih se morfosintaktičkim oznakama (engl. *morphosyntactic descriptors* – MSD). Za oznake koje čuvaju samo osnovnu informaciju o vrsti riječi u literaturi se ustalio izraz POS-oznake (engl. *part of speech tags*) no ovdje će se koristiti izraz *oznake vrste riječi*. U ovom se radu prvenstveno govori o oznakama vrsta riječi jer skup oznaka korištenih u označavanju i pokusima, kako će se to u kasnijim poglavljima pokazati, broji samo 41 različitu oznaku. Rezultati drugih označivača koji se u radu spominju odnose se na rezultate u označavanju vrsta riječi (engl. *POS tagging*) gdje su skupovi oznaka relativno maleni i najčešće sadržavaju do oko 100 oznaka.

U okviru ovog rada razvijeno je programsko ostvarenje automatiziranog označivača vrsta riječi zasnovanog na nadziranom strojnom učenju. Pri tome je korišten skriven Markovljev model, stohastički model vrlo prikladan primjeni na označavanje vrsta riječi. Osim toga, u okviru ovog rada provedeno je i ručno označavanje vrsta riječi u relativno malenom korpusu za potrebe provedbe nadziranog učenja. Korpus sadrži tekst na hrvatskom jeziku, što je u skladu s namjerom ostvarenja automatiziranog označivača vrsta riječi za hrvatski jezik.

U sljedećem poglavlju dan je pregled ranijih radova iz područja, a u poglavlju 3 opisana je problematika označavanja vrsta riječi uz osvrt na potrebne općenite jezične i morfološke temelje. Poglavlje 4 donosi opis rada skrivenog Markovljevog modela, dok su u poglavlju 5 opisani neki detalji konkretnog programskog ostvarenja. Najzanimljiviji je ipak sadržaj poglavlja 6 u kojem se mogu pronaći rezultati testiranja ostvarenog označivača, kao i brojni detalji o korištenom ručno označenom korpusu.

2. Dosadašnji rad na označavanju vrsta riječi

2.1. Rani period razvoja

Istraživanja u području obrade prirodnih jezika počinju već pedesetih godina dvadesetog stoljeća, vrlo rano u povijesti razvoja računalne znanosti. U tom se desetljeću pojavljuju neki bitni radovi koji utiru put budućim istraživanjima. Računalna obrada prirodnog jezika u čvrstoj je sprezi s istraživanjima umjetne inteligencije od samih njihovih početaka. To se može vidjeti i u Turingovoj definiciji ispita inteligencije stroja uvelike zasnovanoj na problemu sinteze i analize prirodnog jezika (Turing, 1950). U pokusu provedenom na sveučilištu Georgetown 1954. godine po prvi put je iskušano potpuno automatizirano prevođenje s jednog na drugi prirodni jezik – prevedeno je šezdeset rečenica s ruskog na engleski jezik, kao što bi bilo primjereno duhu vremena hladnog rata (Hutchins, 2004).

Rad s većim korpusima započinje već šezdesetih godina s jednim od pionirskih djela korpusne lingvistike – objavom knjige *Computational Analysis of Present Day American English* osnovnih statističkih podataka o korpusu prikupljenom na sveučilištu Brown u SAD-u (Kucera i Francis, 1967). Prikupljeni korpus, prikladno nazvan *Brown corpus*, sadržavao je tekstove različitog podrijetla: novinske tekstove, tekstove iz različitih časopisa, državnih dokumenata i izvještaja te knjiga (znanstvenih knjiga i članaka, različitih djela fikcije, romana, novela, eseja i sl.). Sadržavao je otprilike jedan milijun riječi. Korpus Brown važan je i zato što su u njemu sve riječi, tj. značke bile označene oznakama iz skupa od oko 80 definiranih oznaka. Drugi važni korpusi engleskih tekstova su korpus Lancaster-Oslo-Bergen (skraćeno: LOB), po veličini gotovo identičan korpusu Brown i korpus Wall Street Journala (skraćeno: WSJ), jedan od trenutno popularnijih označenih engleskih korpusa. Za označavanje vrsta riječi u engleskom jeziku uobičajen je manji broj oznaka, dok najrazrađeniji skupovi oznaka sadrže do oko 170 različitih oznaka (Zavrel i Daelemans, 1999; Agić i Tadić, 2006).

Za usporedbu, za označavanje vrsta riječi u hrvatskom koriste se skupovi oznaka koji broje 896 oznaka (Agić i Tadić, 2006; Agić et al., 2008), dok u hebrejskom jeziku skupovi oznaka sadrže i preko 3651 oznaka (Goldberg et al., 2008).

2.2. Nadzirano i nenadzirano učenje

Označivač vrsta riječi može učiti nadzirano i nenadzirano. Bitno je naglasiti da je podjela označivača na one koji uče nadzirano i one koje uče nenadzirano ortogonalna na podjelu označivača prema pristupima koje koriste.

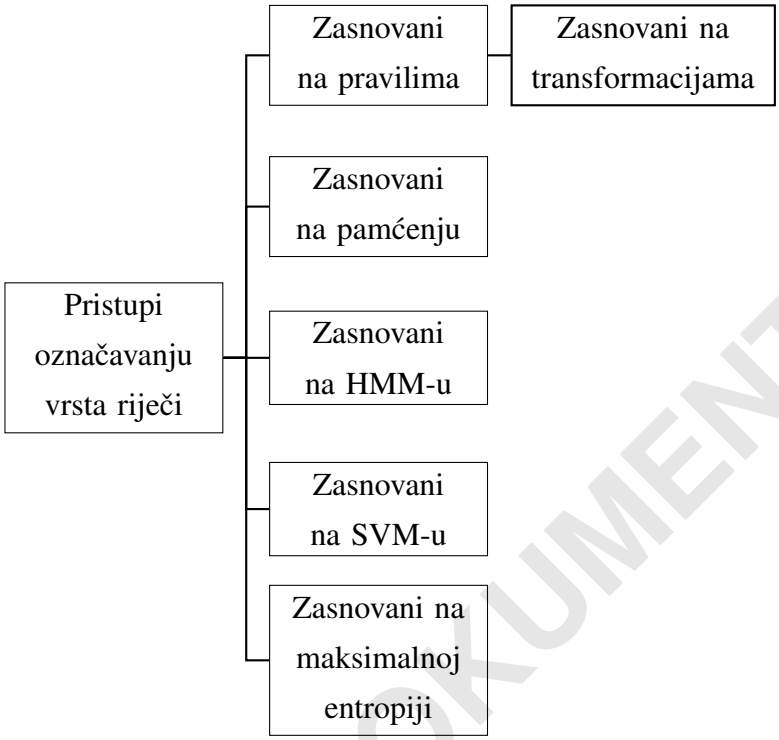
Modeliranje i ostvarenje označivača koji uči nadzirano u pravilu je nešto izravnije i jednostavnije, no pretpostavlja postojanje unaprijed označenog korpusa. Nadzirano učenje provodi se pomoću temeljne istine sadržane u oznakama riječi u tekstu. Ove oznake riječima u tekstu dodjeljuju uvježbani ljudski označivači, često lingvisti. Postupak ručnog označavanja uvijek je dugotrajan posao, pa ga se ponekad izbjegava korištenjem nenadziranog učenja označivača. Nenadzirano učenje ne zahtijeva označeni korpus no zahtijeva izradu algoritama koji iz neoznačenog korpusa mogu sami izvući zaključke o vrsti riječi. Označivači zasnovani na nenadziranom učenju obično dostižu niže razine točnosti označavanja od onih zasnovanih na nadziranom učenju (Goldberg et al., 2008).

2.3. Pristupi označavanju riječi

Pristupi označavanju vrsta riječi mogu se podijeliti na šest skupina:

1. označavanje zasnovano na pravilima (engl. *rule-based*),
2. označavanje zasnovano na transformacijama (engl. *transformation-based*),
3. označavanje zasnovano na pamćenju (engl. *memory-based*),
4. označavanje zasnovano na SVM-u,
5. označavanje zasnovano na maksimalnoj entropiji (engl. *maximum entropy-based*) i
6. označavanje zasnovano na HMM-u.

Možda nije praktično izdvojiti jednu kategoriju stohastičkih označivača, uzevši u obzir činjenicu da se stohastički postupci koriste u više prethodno navedenih skupina poput



Slika 2.1: Taksonomijski prikaz pristupa razvoju označivača vrste riječi

.

onih opisanih u (Ratnaparkhi et al., 1996), (Ratnaparkhi, 1997) ili pak poput onog opisanog u (Brill, 1995). Njihovi međuodnosi i razvojni putevi mogu se vidjeti u taksonomijskom prikazu na slici 2.1.

2.3.1. Označivači zasnovani na pravilima

Označivači zasnovani na pravilima provode označavanje postupkom koji se sastoji od dva dijela. Prvi dio svakoj riječi pridijeli najvjerojatniju oznaku, odabranu na temelju analize velikog korpusa, bez obzira na njezin kontekst, dok drugi eventualno popravlja dodijeljenu oznaku koristeći skup ručno napisanih pravila (Brill, 1992). Iako najčešće ručno pisana, pravila se mogu i strojno naučiti, kao što je to napravljeno u (Hajič i Hladká, 1997)). Sužavanje potencijalnih oznaka za neku riječ s punog skupa oznaka na nekoliko uistinu mogućih obavlja se korištenjem morfološkog analizatora. Primjer morfološkog analizatora često korištenog za engleski jezik je *English morphological analyser* (ENGTWOL). Takvi analizatori za svaku zadanu riječ, neovisno o njezinom kontekstu, ispisuju moguće oznake vrste riječi (Forsythe, 2008). Pa tako za riječ *watches* sustav ENGTWOL ispisuje

"watch" <SV> <SVO> <InfComp> V PRES SG3 VFIN @+FMAINV

i *"watch" N NOM PL*

što, kraće rečeno, govori o tome da bi navedena riječ mogla biti i glagol i imenica.[1]

Malo drugačiji pristup, također sastavljen od dva dijela, prvo svakoj riječi dodjeljuje najvjerojatniju oznaku (primjerice, riječ *među* je najčešće prilog, a rjeđe imenica ili pak neka druga vrsta riječi) te potom izvodi dodatne prolaske kroz tako označen tekst i nad njime primjenjuje pravila za dodjelu oznaka (Brill, 1992). Ručno napisana pravila opisuju slijedove oznaka koji se smiju (ili pak ne smiju) pojaviti u rečenicama. Pravila u ovakvim označivačima, na primjer, imaju oblik poput (Brill, 1992):

TO IN NEXT-TAG AT.

Ovakvo pravilo zapravo izriče sljedeću uputu: "Ako je riječ označena oznakom *TO*, a sljedeća riječ je označena oznakom *AT*, promijeni oznaku iz *TO* u *IN*".[2] Pravilo ima smisla ako se u obzir uzme činjenica da je u engleskom jeziku puno vjerojatnije susresti imeničku frazu nakon prijedloga, a ne infinitivskog *to*.

Postupanje s nepoznatim riječima ne spominje se eksplicitno u (Brill, 1992), no u (Hajič i Hladká, 1997) opisan je postupak u kojem se jednostavnim, fiksiranim pravilima svim riječima koje završavaju nekim unaprijed zadanim sufiksom dodjeljuje zadana oznaka.

Označivač opisan u (Brill, 1992) testiran je na manjem dijelu (5%) korpusa Brown i ostvario je točnost od 95%. Primjer označivača zasnovanim na pravilima je i sustav TAGGIT koji je s točnošću od 77% pomagao u označavanju korpusa Brown (Greene i Rubin, 1971), kao i neki drugi sustavi čija točnost nije objavljena ili je neusporediva s drugim sustavima zbog netipičnog mjerenja točnosti (Koskenniemi, 1990). Ispitivanja označivača iz (Brill, 1992), doduše modificiranog, provedena su i za češki jezik te je postignuta točnost označavanja od 79,75% (Hajič i Hladká, 1997).

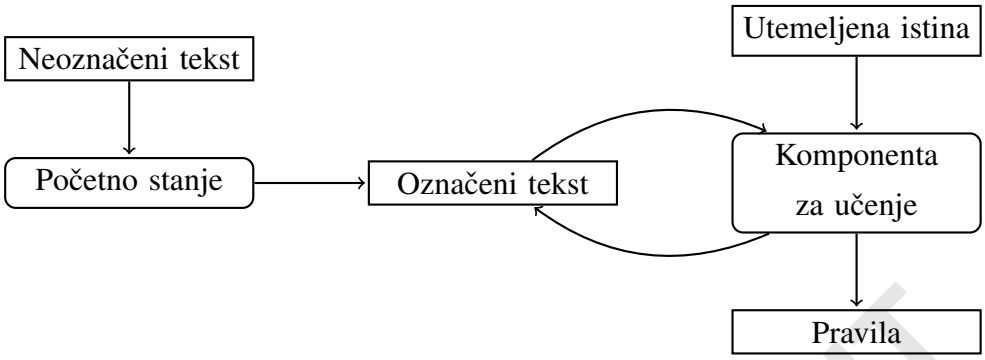
2.3.2. Označivači zasnovani na transformacijama

Početni rad na označavanju zasnovanom na transformacijama započeo je sredinom 1990-ih godina, u vrijeme kada su stohastički označivači (često zasnovani na skrivenim Markovljevim modelima[3]) bili popularan alat u tada objavljivanim člancima.

[1] Primjer preuzet s internetskog sučelja sustava ENGTWOL koji se može pronaći na adresi http://www2.lingsoft.fi/cgi-bin/engtwol

[2] TO = infinitivski *to* (engl. *infinitive to*), AT = član (engl. *article*), IN = prijedlog (engl. *preposition*)

[3] engl. *hidden Markov model*, skraćeno HMM



Slika 2.2: Pregled principa rada označivača zasnovanog na transformacijama (preuzeto i prilagođeno iz (Brill, 1994))

Primjedba koja se mogla uputiti (vrlo uspješnim) označivačima zasnovanim na HMM-ima je njihovo indirektno pohranjivanje lingvističkih informacija u velikim statističkim tablicama koje sadrže desetke tisuća kontekstnih i leksičkih vjerojatnosti. Vrlo je prirodna i logična bila želja zapisati, "uhvatiti" relevantne lingvističke informacije u manjem broju jednostavnih nestohastičkih pravila. Uputno je stoga bilo ponovno razmotriti pristup zasnovan na pravilima, uz neke dodane modifikacije. Razlika u odnosu na obične sustave zasnovane na pravilima manifestira se u uvođenju nove metode učenja označivača. Učenje zasnovano na transformacijama, vođeno smanjenjem greške (engl. *transformation-based error-driven learning*) je metoda koja je uvedena i korištena u istraživanjima opisanim u (Brill, 1994, 1995). Učenje zasnovano na transformacijama, vođeno smanjenjem greške razradio je sredinom 1990-ih Eric Brill (Brill, 1994). Postupak takvog učenja sastoji se od nekoliko faza. Jezgrovit prikaz postupaka može se vidjeti u dijagramu na slici 2.2, no može ga se ukratko opisati u sljedećih nekoliko točaka:

- neoznačeni tekst (tekst A) prvo se označi početnim označivačem (to može biti razrađen označivač, no dovoljan je i označivač koji dodjeljuje slučajnu oznaku svakoj znački),
- tekst označen početnim označivačem uspoređuje se s istinom (engl. *truth*), tj. sa strukturama kakve se pojavljuju u drugom skupu tekstova, skupu koji je bio ručno označen (nazovimo taj skup tekst B),
- pri tome se uče transformacije koje mijenjaju pravila s kojima je početni označivač označio značke u tekstu A u pravila s kojima bi označavanje teksta A više nalikovalo označenom tekstu B,
- te se konačno postupak iterativno ponavlja pri čemu se pohlepno odabiru tran-

sformacije s kojima se ostvaruju najbolje točnosti označavanja.

U osnovi, sustav na označenoj rečenici (u postupku učenja uzimaju se, dakako, sve rečenice u dostupnom korpusu) iskušava sve predloške za pravila poput onih opisanih u prethodnom potpoglavlju, u njima varira sve oznake i odabire transformaciju, tj. pravilo koje rezultira najvećim smanjenjem pogreške u označavanju. Učenje završava kada se smanjenje greške padne ispod unaprijed zadanog praga (Brill, 1994).

Označavanje nepoznatih riječi obavlja se pomoću metode koja se oslanja na informaciju sadržanu u prefiksima i sufiksima riječi. U početnom označavanju nepoznatim se riječima dodjeljuje oznaka *vlastita imenica* ako počinju velikim slovom ili *opća imenica* u suprotnom slučaju (Brill, 1994). Nakon toga na nepoznate se riječi primjenjuju pravila koja imaju oblik poput[4]

"Ako su prva četiri znaka riječi jednaka Z, promijeni oznaku iz X u Y"

ili pak

"Ako oduzimanjem (dodavanjem) nekoliko znakova s početka ili kraja riječi dobivamo poznatu riječ, promijeni oznaku iz X u Y".

Ovakvim označavanjem na nepoznatim se riječima postižu točnosti do 82,2% (Brill, 1995). Testovi provedeni na korpusu WSJ, čijih je 950000 znački iskorišteno za treniranje (od čega je 600000 korišteno za učenje kontekstnih pravila a 350000 za učenje pravila za označavanje nepoznatih riječi) a 150000 znački za testiranje, postiže ukupnu točnost od 96,6% (Brill, 1995).

Kao što to pokazuju gore navedeni (zadnji objavljeni) članci, već sredinom 1990-ih godina zamire interes za istraživanjem označavanja vrsta riječi zasnovanog na pravilima i transformacijama.

2.3.3. Označivači zasnovani na pamćenju

Označivači zasnovani na memoriji tako se nazivaju zato što je u srži njihovog rada učenje zasnovano na pamćenju,[5] vrsta nadziranog učenja koje se temelji zaključivanju prema sličnosti (engl. *similarity-based reasoning*). Učenje zasnovano na pamćenju oblik je nadziranog induktivnog učenja iz danih primjera. Primjeri su predstavljeni vektorima vrijednosti različitih njihovih značajki. Tijekom učenja, primjeri se redom

[4] Ovi prilagođeni primjeri također su preuzeti iz (Brill, 1994).

[5] Engl. *memory-based learning*; koriste se i izrazi *instance-based, example-based, exemplar-based, case-based, analogical, lazy* koji označavaju istu vrstu učenja (Daelemans et al., 1996).

prikazuju klasifikatoru koji ih tada sprema u memoriju. Tijekom testiranja se klasifikatoru zadaje skup novih, neviđenih primjera te se za svaki od tih novih primjera računa njegova udaljenost od svih zapamćenih primjera. Novom se neviđenom primjeru tada pripisuje kategorija kojoj pripada njemu najbliži zapamćeni primjer (ili skupina primjera). Popularan klasifikator koji implementira učenje zasnovano na pamćenju poznati je k-NN, klasifikator koji razmatra k najbližih susjeda.

Ovu se teorijsku pozadinu u sferu označavanja najlakšemože preslikati ako se razmotri mjera udaljenosti između primjera. Uobičajena mjera u ovakvim klasifikatorima izražena je formulom

$$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{2.1}$$

gdje se udaljenost između dviju vrijednosti definira kao

$$\delta(x_i, y_i) = \begin{cases} 0, & \text{ako } x_i = y_i \\ 1, & \text{inače} \end{cases} \tag{2.2}$$

Ovakva diskretna metrika, opisana u (Daelemans et al., 1996), koristi se zato što su sve značajke primjera simboličke što prirodno nameće primjena na oznake vrsta riječi. Svaka riječ, zajedno sa svojom oznakom, kao i kontekstom (riječima i njihovim oznakama) koji se pojavljuje prije i poslije riječi, jedan je primjer u klasifikatoru. Značajke primjera nisu zapisane kao vektor realnih brojeva, već kao niz oznaka. Ako se uzme u obzir da je, primjerice, za oznaku riječi u fokusu važnija ona sama od riječi koje se pojavljuju u njenom lijevom i desnom kontekstu, dolazi se do potrebe za uvođenjem težina pojedinih značajki. To se može ostvariti množenjem svake značajke s njezinom informacijskom dobiti – brojem koji izražava smanjenje entropije skupa za učenje, jednom kada je poznata vrijednost te značajke, kao što to opisuje formula iz (Daelemans et al., 1996)

$$\Delta(X, Y) = \sum_{i=1}^{n} G(f_i)\delta(x_i, y_i). \tag{2.3}$$

Prednost označivača zasnovanog na pamćenju je to što on ne zahtijeva komponentu koja provodi dodatno zaglađivanje, što je obavezan dio stohastičkih označivača ako su podaci koji se koriste za učenje rijetki. Osim toga, u takvom označivaču čak i uzorci koji se ne pojavljuju često mogu doprinijeti generalizaciji klasifikatora (Zavrel i Daelemans, 1999). Nedostaci svih metoda zasnovanih na učenju temeljenom na pamćenju su njihovo odgađanje svog računanja do trenutka kada treba klasificirati novi primjer, kao i velike količina memorije potrebne za pohranu svih "naučenih" primjera. Ovom se drugom nedostatku može doskočiti korištenjem strukture podataka koja uvelike smanjuje količinu potrebne memorije i ubrzava dohvat pohranjenih primjera, kao što su to

učinili autori u (Daelemans et al., 1996) uporabivši IGTrees, heurističku aproksimaciju algoritma IB-IG.

Problem označavanja nepoznatih riječi rješava se tako što im se kao značajke odabire po jedna riječ iz njihovog lijevog i desnog konteksta, uz prvo slovo prefiksa i zadnja tri slova sufiksa.

Točnosti dobivene ovim pristupom iznose, ovisno o korpusu nad kojim su testiranja obavljena, između 96,4% i 97,0%, uz točnost na nepoznatim riječima u intervalu između 81,0% i 90,0%.

2.3.4. Označivači zasnovani na SVM-u

SVM (engl. *support vector machines*) je klasifikator opće namjene, prvi put opisan u (Cortes i Vapnik, 1995) koji, primijenjen na problem označavanja vrsta riječi, ostvaruje rezultate koji su po svojoj točnosti pri vrhu popisa rezultata različitih označivača. Osim toga, pokazuje se da ga se vrlo uspješno može primijeniti i na problem detekcije pogrešaka u ručno označenim korpusima (Nakagawa i Matsumoto, 2002). Klasifikator SVM uči nadzirano, tj. zahtijeva korpus s označenim vrstama riječi. Temeljna ideja klasifikatora SVM jest odvojiti primjere za učenje (modelirane kao višedimenzijske vektore), u slučaju s dva razreda, u dvije skupine tako što se između tih dviju skupina konstruira (*„uči “*) hiperravnina pozicionirana tako da osigura maksimalan odmak između same hiperravnine i svake od dviju grupa primjeraka za učenje. Detaljniji opis SVM-a nadilazi opseg ovog rada, no potpunije objašnjenje njegovog rada može se pronaći u (Cortes i Vapnik, 1995; Drucker et al., 1997).

Problem preslikavanja na specifični problem, u ovom slučaju označavanje vrsta riječi, u najvećoj se mjeri, kao što je to obično slučaj u primjeni klasifikatora opće namjene, manifestira u odabiru značajki kojima će se modelirati svaki primjer koji će se koristiti za učenje ili testiranje. Kada se radi o označavanju vrsta riječi, za značajke primjera (a primjer je zapravo svaka riječ koja se pojavljuje u korpusu) odabiru se različiti podnizovi znakova te riječi, kao i njezine okoline – leksička okolina (riječi koje se pojavljuju lijevo i desno od promatrane riječi) i tzv. kontekstna okolina (koja uključuje oznake tih riječi koje se pojavljuju lijevo i desno od promatrane) (Nakagawa et al., 2001). Sve su navedene značajke binarne i izgledaju poput onih opisanih u (Giménez i Màrquez, 2003):

prethodna_riječ_je_the

ili pak

prethodne_dvije_oznake_su_DT_NN.

Značajke kojima se opisuju nepoznate riječi odabrane su slično kao i u drugim radovima poput (Brill, 1995); koriste se:

kontekst oznaka riječi: oznake po dviju riječi s lijeve i s desne strane promatrane riječi,

leksički kontekst (kontekst u smislu susjednih riječi): leksički oblici po dviju riječi s lijeve i desne strane promatrane riječi i

podnizovi znakova: prefiksi i sufiksi nepoznate riječi duljine maksimalno četiri znaka kao i postojanje brojki, crtica, svih velikih slova ili pak samo početnog velikog slova u nepoznatoj riječi.

Rezultati označivača koji koriste SVM nisu se znatno promijenili od prvih objavljenih radova; obično se pozicioniraju u interval između 97,10% i 97,16%, s točnostima za nepoznate riječi između 83,5% i 88,5% (Nakagawa et al., 2001; Giménez i Marquez, 2004). Korisno je napomenuti činjenicu da je SVMTool, alat čiji je prototip razvijen u (Giménez i Màrquez, 2003), a brža verzija, implementirana u jeziku C++, objavljena kasnije na internetskim stranicama Politehničkog sveučilišta u Kataloniji,[6] slobodno dostupan istraživačima, što se za veliki broj označivača razvijanih u okviru drugih znanstvenih članaka ne može reći.

2.3.5. Označivači zasnovani na maksimalnoj entropiji

U temeljima pristupa označavanju vrsta riječi zasnovanog na maksimalnoj entropiji utkane su posljedice poznatog postulata o maksimalnoj entropiji, izrečenog u (Jaynes, 1957; Good, 1963), koji se može izreći na sljedeći način:

> *Uz zadani skup podataka i zadani skup ograničenja (provjerljivih informacija), razdioba vjerojatnosti koja najbolje opisuje takav skup je upravo ona koja koja ima najveću entropiju.*

Ovakva apstraktna definicija ne govori puno pa zato valja dobro razmotriti kako se problem označavanja vrsta riječi rješava ovakvim pristupom. Označavanje zasnovano na maksimalnoj entropiji inherentno je blisko vezano za druge pristupe temeljene na korištenju statističkih podataka iz korpusa i vjerojatnosti pojave određenih uzoraka u njemu. Vjerojatnosni model definira se nad skupom $H \times T$, u kojem je s H označen

[6] Alat se može preuzeti na stranici http://www.lsi.upc.es/ nlp/SVMTool/.

skup svih leksičkih konteksta i konteksta u vidu oznaka neke promatrane riječi, a T označava skup svih dopuštenih oznaka vrsta riječi (Ratnaparkhi et al., 1996). Svaki element h skupa H naziva se poviješću (engl. *history*) neke riječi (Ratnaparkhi et al., 1996). Vjerojatnost da se nekoj riječi, uz njezinu povijest h, pripiše oznaka t definira na sljedeći način:

$$p(h,t) = \pi\mu \sum_{j=1}^{k} \alpha_j^{f_j(h,t)}, \tag{2.4}$$

pri čemu se s k označava broj značajki. Značajke povijesti $f_j(h,t)$ poprimaju vrijednosti iz skupa $\{0,1\}$, efektivno govoreći sadrži li ili ne sadrži povijest promatrane riječi neku karakteristiku. Oznake π, μ i α_i označavaju redom normirajuću konstantu i pozitivni parametri modela. Vrijednosti ovih parametara odabiru se tako da se pomoću vjerojatnosti p maksimira izglednost skupa za učenje (Ratnaparkhi et al., 1996):

$$L(P) = \prod_{i=1}^{n} p(h_i,t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}. \tag{2.5}$$

U prethodnoj formuli s n je označen broj riječi u nizu koji treba označiti. Ovako izražen model može se reinterpretirati formalizmima maksimalne entropije, gdje je cilj maksimirati entropiju razdiobe vjerojatnosti p definirane kao:

$$H(p) = -\sum_{h\in H, t\in T} p(h,t) log[p(h,t)]. \tag{2.6}$$

Opširan opis svih algoritama koji se koriste u postupku učenja ovakvog označivača, kakav se može pronaći u (Ratnaparkhi et al., 1996; Ratnaparkhi, 1997; Toutanova i Manning, 2000; Toutanova et al., 2003), ovdje je preskočen, no može se spomenuti da se nepoznate riječi opet tretira na sličan način kao i drugdje, uzimajući u obzir informaciju koju nose njezini sufiksi, velika slova koja se u riječi pojavljuju (početno slovo ili sva slova) te postojanje crtica u riječi. Učenje označivača obavljeno je na oko milijun znački korpusa WSJ sa skupom oznaka koji propisuje Penn Treebank, dok je za testiranje korišten manji skup od oko 130000 znački. Označivači zasnovani na maksimalnoj entropiji dosežu točnosti od 85,56% na nepoznatim riječima i 96,63% ukupno, kakve su dobivene u (Ratnaparkhi et al., 1996), do 89,04% za nepoznate i 97,24% ukupno, što je dobiveno u jednom novijem radu (Toutanova et al., 2003).

2.3.6. Označivači zasnovani na HMM-u

Skriveni Markovljevi modeli vrlo su rano, već krajem 1980.-ih godina, stekli popularnost u istraživanju označavanja vrsta riječi. Neki od prvih sustava za rješavanje

problema višeznačnosti riječi, odnosno problema odabira vrste kojoj neka promatrana riječ pripada, koriste stohastičke pristupe čija se forma uvelike preklapa s formom kakvu propisuje HMM (Church, 1988). Primjerice, u (DeRose, 1988) opisuje se algoritam dinamičkog programiranja koji odabire oznake vrste riječi za svaku riječ u rečenici tako što optimira umnožak vjerojatnosti da se riječi i pridijeli oznaka vrste j i vjerojatnosti da se prije riječi označene s j pojavi riječ označena s k. Iako se ne koristi uobičajena notacija i prihvaćeni naziv, navedena formulacija ustvari opisuje Viterbijev algoritam koji je šire prihvaćanje pronašao nakon objave Rabinerovog često citiranog (iako ne i prvog) rada kojim su sustavno opisani skriveni Markovljevi modeli i algoritmi koji se nad njime provode (Rabiner, 1989). Uporaba HMM-a na konkretne probleme opisivana je i ranije, primjerice u (Levinson et al., 1983).

HMM je vjerojatno najčešće korišten matematički model za označavanje vrsta riječi te se kroz veći broj objavljenih radova pokazao kao robustan, efikasan, točan, prilagodljiv i ponovno iskoristiv (misli se na njegovu primjerenost različitim korpusima i različitim jezicima) alat (Cutting et al., 1992). Rezultati prvih označivača zasnovanih na HMM-u, iako ne opisuju izričito rezultate postignute nad nepoznatim riječima, prilično su visoki – dosežu razine od oko 96%, a na nekim korpusima spominju i točnosti od 97%, postizane u (Cutting et al., 1992; Merialdo, 1994). Nedostatak HMM-a je problem rijetkih podataka (engl. *data sparseness*). Nizovi riječi ili njihovih oznaka koji se ne pojave u skupu za učenje imaju u HMM-u vjerojatnost nula. U strukturama podataka koje čuvaju vjerojatnosti postoji veliki broj nul-vrijednosti koje znatno utječu na krajnje rezultate označivača. Zbog toga se nužno rabe metode zaglađivanja vjerojatnosti, odnosno postupci kojima se vjerojatnostima koje imaju vrijednost nula dodjeljuje jedna procijenjena malena vrijednost veća od nule, kao u (Zavrel i Daelemans, 1999).

U okviru ovog diplomskog rada u cijelosti je razvijen vlastiti označivač zasnovan na HMM-u pa se tako potpun opis svih detalja rada takvog označivača može pronaći u poglavlju 5.

2.4. Pregled radova prema jezicima

Rezultati u prethodnom potpoglavlju uglavnom su postizani u ispitivanjima na korpusima na engleskom jeziku pa se u ovom potpoglavlju na jednom mjestu, neovisno o korištenom pristupu, mogu pronaći rezultati postizani na različitim jezicima. Za svaki je rad navedena postignuta točnost, pristup koji je korišten, korpus i referenca na rad u kojem se pobliže može upoznati s izvršenim istraživanjem. Zbog nepostojanja ko-

Jezik	Toč.	Pristup	Korpus	Autori
Češki	93,6	pamćenje	UJOP[7]	(Zavrel i Daelemans, 1999)
Engleski	97,2	maks. entropija	WSJ	(Toutanova et al., 2003)
	86,8	HMM[8]	WSJ	(Goldwater i Griffiths, 2007)
Hebrejski	92,4	HMM[9]	Hebrew Treebank[10]	(Goldberg et al., 2008)
Hrvatski	96,3	HMM	CW100	(Agić et al., 2009)
Nizozemski	95,7	pamćenje	Eindhoven	(Zavrel i Daelemans, 1999)
	97,5	SVM	CGN	(Poel et al., 2007)
Slovenski	97,6	HMM	SVEZ-IJS	(Erjavec i Sárossy, 2006)
Španjolski	97,8	pamćenje	CRATER	(Zavrel i Daelemans, 1999)
Švedski	95,6	pamćenje	SUC	(Zavrel i Daelemans, 1999)

Tablica 2.1: Rezultati označivača na raznim jezicima

nvencije, dijelovi korpusa koje se izdvaja za učenje i testiranje odabiru se na način koji se razlikuje od rada do rada. Zbog sažetosti se u tablici 2.1 ne prikazuju svi detalji o korpusu ili dijelu korpusa korištenom za pojedina istraživanja, nego samo njegovo ime. Također, točnosti (u stupcu označenom kraticom *Toč.*) su izražene u postocima.

2.4.1. Osvrt na radove za hrvatski jezik

Velika većina radova koristila je neki oblik nadziranog učenja. No u praksi je nepostojanje ručno označenog korpusa vrlo česta pojava, pogotovo kada postoji želja za provođenjem istraživanja nad jezicima koji nemaju veliki broj govornika kao što je to slučaj s hrvatskim. Pokazuje se da u puno slučajeva ovakvi jezici nažalost često nemaju označene jezične resurse. Zbog ovog razloga razvijani su označivači koji uče nenadzirano. U tablici 2.1 može se vidjeti da su rezultati nenadzirano naučenih označivača uvijek nešto lošiji od rezultata onih koji uče nadzirano. Zbog nepostojanja označenog hrvatskog korpusa slobodno dostupnog za akademska istraživanja, razmatrana je opcija ostvarenja ovakvog označivača, no umjesto toga je, uz trud kakav to

[7] Korpus Instituta za češki jezik u pragu (češ. *Ústav jazykové a odborné přípravy Univerzity Karlovy v Praze*).

[8] Autori su koristili tzv. potpuni bayesovski pristup nenadziranom učenju HMM-a.

[9] Autori su koristili nenadzirano učenje.

[10] Sadrži novinske materijale lista Haaretz, a korišteni su i transkripti iz Knesseta.

uvijek iziskuje, ručno označen manji skup tekstova s ukupno 20000 znački. Veličina označenog korpusa korištenog za učenje i testiranje označivača razvijenog u okviru ovog diplomskog rada primjerenija je uporabi za testiranje označivača koji bi koristio nenadzirano učenje. Također treba spomenuti i činjenicu da se gotovo svi rezultati u stvarnosti osciliraju u ovisnosti o korištenom dijelu korpusa za učenje, skupa oznaka i sličnih okolnosti u kojma su provedena mjerenja točnosti označavanja.

Rezultate dobivene označavanjem u tekstovima na hrvatskom jeziku preporučljivo je uspoređivati s rezultatima dobivenim na sličnim jezicima. U skupinu jezika sličnih hrvatskom mogu se ubrojiti mnogi slavenski jezici, jezici koji s hrvatskim dijele breme, no i bogati skriveni informacijski potencijal, visoke fleksije, tj. složene morfologije. Stoga je, dok nisu izvršena prva mjerenja za hrvatski jezik (Agić i Tadić, 2006), bilo zahvalno proučavati rezultate kakvi su postizani na jezicima sličnih karakteristika, no s bogatijim fondom objavljenih članaka.

3. Problem označavanja vrsta riječi

3.1. Morfologija i vrste riječi

3.1.1. Osnove morfologije

Značenjski sadržaj tekstu se najvećim dijelom pridijeljuje odabirom riječi koje označavaju ciljane teme, radnje i slične značajke. Drugim riječima, značenje koje je inherentno odabranim riječima prenosi se na višu razinu i tvori značenje skupine riječi, rečenice, odlomka i, konačno, cijelog teksta. Pritom je nezanemariv dio značenja sadržan u različitim oblicima iste riječi i u konstrukcijama ostvarenim postavljanjem riječi u razne međusobne odnose. Ovaj se rad bavi vrstama riječi pa se stoga u ovom poglavlju razmatraju oblikovne, tj. morfološke značajke riječi u hrvatskom jeziku i njihov utjecaj na semantiku. Kratko razmatranje jezične strane problema nužno zahtijeva ulazak u sferu lingvistike. Od pomoći će biti sljedećih nekoliko definicija, među njima i definicija morfologije (Matthews, 2009):

> *Morfologija je grana lingvistike koja se bavi različitim oblicima riječi i unutarnjom strukturom riječi u različitim uporabama i konstrukcijama.*

Pri tome se morfem definira kao najmanji jezični odsječak riječi koji ima vlastito značenje. Morfemi mogu biti

1. *slobodni*, kao, na primjer, morfem *jak* koji ujedno sam čini cijelu riječ, ili pak
2. *vezani*, poput morfema *pre* u riječi *prejak*.

Isti će se morfem pojavljivati kao dio razlitih riječi, pa se, prema tome, ovisno o svom morfološkom okruženju, može pojaviti u više inačica. Primjerice, morfem *pod* pojavljuje se i kao dio riječi *podnajam* i kao dio riječi *potpisnik*. Obje njegove inačice nose isto značenje, no u obliku im se vidi malena, glasovnom promjenom uzrokovana razlika. Ovakve inačice koje služe kao izraz istoga morfema nazivaju se alomorfi (Barić et al., 2005). Jezici se prema svojim morfološkim svojstvima mogu raspodijeliti u jedan od tri razreda (Radford et al., 2009):

1. izolirajući jezici (engl. *isolating languages*),

2. sljepljujući jezici (engl. *agglutinating languages*) i

3. flektivni jezici (engl. *inflectional languages*).

Kako to često biva u znanosti, ova podjela jezika po njihovim morfološkim svojstvima nije jedina moguća. Bitno je naglasiti i činjenicu da razgraničenja između ove tri vrste jezika često nisu potpuno oštra. Stoga valja uvijek na umu imati pojam o neizrazitim granicama između ovih vrsta jezika (neki su jezici npr. i sljepljujući i flektivni).

Primjeri izolirajućih jezika su kineski, vijetnamski i niz drugih dalekoistočnih jezika, kao i neki zapadnoafrički jezici. Engleski jezik se često opisuje kao u velikoj mjeri analitički jezik pri čemu se riječ *analitički*[1] najčešće koristi kao sinonim za izolirajući. Idealan izolirajući jezik odlikuje odsutstvo vezanih morfema. Drugim riječima, za razliku od hrvatskog jezika gdje bismo koristeći riječi *pisati* i vezani morfem *-ar* stvorili imenicu *pisar*, u savršenom izolirajućem jeziku koristimo dvije izolirane riječi (npr. *pisati* i *osoba*) povezane u složenicu strukture koja ugrubo izgleda kao *pisati* + *osoba* (Radford et al., 2009). Drugačije rečeno, u ekstremno izolirajućim jezicima svaka se riječ sastoji od točno jednog morfema; riječi se ne modificiraju fleksijom.

U sljepljujućim se jezicima riječi oblikuju korištenjem velikog broja morfema. Primjeri sljepljujućih jezika su, između ostalih, turski, baskijski, gruzijski te donekle i japanski i korejski jezik.

Infleksijski jezici (ponekad ih se naziva i fuzijskim jezicima) poput grčkog i latinskog su jezici u kojima se riječi najčešće oblikuju slaganjem morfema na takav način da je dobivenu riječ teško segmentirati (Matthews, 2009). U lingvističkoj terminologiji riječ *fleksija* (engl. *inflection*) označava općenitu promjenu riječi u svrhu iskazivanja različitih gramatičkih kategorija poput glagolskog vremena, roda, broja i sl. *Deklinacija* (engl. *declension*) je posebna vrsta fleksije koja se primjenjuje na, primjerice, imenice i pridjeve, dok je *konjugacija* (engl. *conjugation*) posebna vrsta fleksije koja se primjenjuje na glagole. U flektivnim jezicima skoro svaka riječ, iako zasebna cjelina, može poprimiti niz različitih značenjskih uloga. Time se primjerice jedan glagol flektivno mijenja kako bi izrazio vrijeme radnje (prezent, perfekt, futur ili drugo), vid (svršeni ili nesvršeni), stanje (aktiv ili pasiv) i sl. Za primjer kojim se značajke flektivnih jezika mogu okvirno opisati može se uzeti riječ *bonus* iz latinskog jezika. To je ujedno i osnovni oblik (muški rod jednine u nominativu), tj. lema te riječi. Kako bi se promijenila neka od karakteristika riječi (rod, broj ili padež), potrebno je nastavak *-us*

[1] Iako ne u svim, u većini konteksta su riječi *izolirajući* i *analitički* sinonimi (Matthews, 2009).

zamijeniti drugim nastavkom, primjerice s *-um* koji tada može značiti i akuzativ muškog roda i akuzativ srednjeg roda i nominativ srednjeg roda. Kod nekih se primjera flektivnih jezika, npr. neki romanski (npr. španjolski) i neki germanski jezici (npr. njemački), njihova flektivnost ponajviše manifestira u glagolima, dok se u drugima, poput slavenskih jezika (češki, poljski, ruski, slovenski, hrvatski i drugi), bogata flektivnost pojavljuje u svim promjenjivim vrstama riječi.

3.1.2. Vrste riječi

U gotovo svim svjetskim jezicima riječi se dijele na oko deset osnovnih vrsta. One su nabrojane i pobliže opisane u nastavku teksta. U literaturi se kao sinonim za vrstu riječi mogu pronaći i izrazi *razred riječi*, *leksički razred* i *leksička kategorija*.[2]

Riječi se u hrvatskom jeziku mogu pronaći u različitim odnosima s drugim riječima u rečenicama. U tom slaganju u rečenice neke riječi trpe morfološke promjene dok druge uvijek ostaju iste. Prema tome riječi u hrvatskom jeziku možemo podijeliti na dvije vrste po promjenjivosti (Raguž, 1997):

- promjenjive i
- nepromjenjive.

U hrvatskom jeziku poznajemo tri sustava promjena riječi (Raguž, 1997):

sklonidbu (deklinaciju): koja se primjenjuje na imenice, zamjenice pridjeve i brojeve,

stupnjevanje (komparaciju): koje primjenjujemo na većinu pridjeva i neke priloge, te

sprezanje (konjugaciju): koje primjenjujemo samo na glagole.

U hrvatskom jeziku postoji deset vrsta riječi. Njih šest, konkretno imenice, pridjevi, brojevi, zamjenice, glagoli i prilozi uvrštavaju se u skupinu promjenjivih vrsta riječi dok se preostale četiri, prijedlozi, veznici, čestice i uzvici uvršavaju u skupinu nepromjenjivih vrsta riječi (Barić et al., 2005). Unatoč tome što se nad nekim prilozima može provoditi komparacija, u literaturi ih se, pogotovo u osnovnoškolskim ili čak srednjoškolskim udžbenicima, ipak često može pronaći uvrštene u skupinu nepromjenjivih riječi (Ham, 2002).

[2] U engleskoj literaturi se za vrste riječi koriste izrazi *part of speech*, *word class*, *lexical class* i *lexical category*.

3.1.3. Otvoreni i zatvoreni skup riječi i višeznačnosti u označavanju

Na primjeru engleskog jezika može se prikazati razlika između starije, tradicionalne podjele riječi na vrste i novije, funkcionalne. Tradicionalna podjela nalaže da se riječi engleskog jezika mogu svrstati u jednu od osam kategorija: imenice, zamjenice, pridjeve, glagole, priloge, prijedloge, veznike ili usklike. Novije podjele uočavaju potrebu za nekim novim kategorijama i neke kategorije dijele na više podkategorija. Osim toga, riječi dijele u jednu od dvije skupine:

1. otvorene riječi i
2. zatvorene riječi.

Zatvorene riječi su skup riječi nekog jezika u koji se nove riječi u pravilu ne mogu dodavati. Kategorija zatvorenih riječi obično sadržava relativno malen dio riječi iz nekog jezika. Tipično se radi o prijedlozima, veznicima, česticama i sličnim vrstama riječi. Funkcionalna podjela vrsta riječi engleskog jezika u zatvorene riječi ubraja primjerice određene i neodređene članove, veznike, prijedloge, čestice, skraćene oblike riječi (*I'm* umjesto *I am*), klitike, glavne brojeve, zamjenice i pomoćne glagole.

Otvorene riječi su skup riječi nekog jezika u koji se relativno često dodavaju nove riječi. Nove riječi mogu se dobiti na različite načine, između ostalih, posuđivanjem iz drugih jezika, srastanjem dviju starih riječi, tj. stvaranjem složenica ili se, kako se to događa u vječno promjenjivim entitetima poput jezika, nova riječ može skovati iz naizgled ničega.

Podjela na otvorene i zatvorene riječi ne spominje se u preporučenoj gramatici hrvatskog jezika (Barić et al., 2005) iako može biti korisna u označavanju vrsta riječi, kako automatiziranom, tako i ručnom. Ako su ljudskom ili računalnom označivaču poznate sve zatvorene riječi, on će znati da ni za jednu riječ koja nije među njima ne smije razmotriti opciju da ju označi, recimo, oznakom neke od vrsta riječi iz otvorene kategorije. Za zatvorene riječi znaju se vrste riječi kojima one pripadaju, bila to jednoznačno određena jedna vrsta riječi (npr. riječ *danas* je uvijek prilog) ili jedna od više mogućih (npr. riječ *poslije* može biti i prilog i prijedlog). Kada se u učenju računalnog označivača riječi koriste veći korpusi, opravdano je pretpostaviti da će označivač naučiti sve riječi iz zatvorene kategorije. Kasnije, pri označavanju nepoznatih riječi, može se s pravom pretpostaviti da je nepoznata riječ iz skupa otvorenih riječi pa se time uvelike sužava broj oznaka koje bi ta riječ mogla poprimiti.

U hrvatskom jeziku postoji veliki broj riječi, pogotovo onih zatvorenih, koje mogu spadati u više od jedne vrste riječi. Dakle, ista riječ može, ovisno o svom kontekstu, biti različite vrste. No vrlo se često te razlike u kontekstima teško uočavaju. Hrvatski jezični portal – HJP[3] tako navodi da riječ *kao* može biti i prilog i veznik. HJP tako dalje navodi da je spomenuta riječ prilog kada, na primjer, označava sličnost ili jednakost pojmova ili postupaka, dok je veznik u situacijama kada izražava približnost ili prividnost. Kako bi se za riječ moglo tvrditi da je određene vrste, ona se mora nalaziti u određenom kontekstu. No ako su opisani zahtjevi na kontekst riječi slični kao u gore navedenom primjeru, čak će i ljudski označivač imati poteškoće u njihovom razlikovanju. Nameće se ideja o uvođenju konvencije koja bi propisivala da se svaku pojavu te riječi označava uvijek istom oznakom. Ovakva je konvencija uvedena u skupu oznaka opisanom u dodatku C.

3.2. Oznake vrsta riječi

Kako bi se označavanje riječi moglo uspješno provesti, potrebno je prvo imati jasno definiran skup oznaka. Kao što je objašnjeno u uvodu, u ovom se radu koristi skup oznaka koji sadrži više od deset najosnovnijih oznaka vrsta riječi, no nije detaljno razrađen poput skupa morfosintaktičkih opisnika kakav propisuje norma MULTEXT-East i kakav bi bilo idealno koristiti kada se radi o jeziku poput hrvatskog. Nažalost, korpus u kojem su riječi označene punim MSD-ima nije bio dostupan za potrebe ovog diplomskog rada pa je jedan manji skup rečenica ručno označen. Zbog vremenske zahtjevnosti ručnog označavanja vrsta riječi u tekstu, koje je obavio autor ovih redaka, odabran je manji skup oznaka vrsta riječi. Označavanje punim MSD-ima zadatak je znatno veće složenosti, a time i veće vremenske zahtjevnosti te bi označavanje teksta punim MSD-ima nadilazilo okvire rada. Kako bi se ilustrirala detaljnost MSD-ova, u nastavku teksta može se, prije prelaska na skup oznaka stvarno korišten u ovom radu, pročitati opis norme MULTEXT-East.

3.2.1. Norma MULTEXT-East

MULTEXT (*Multilingual Text Tools and Corpora*) međunarodni je projekt financiran iz fondova Europske unije namijenjen razvoju višejezičnih korpusa, konvencija njihovog označavanja i upotrebljivih alata za njihovu analizu. Projekt je bio ostvaren u

[3]Hrvatski jezični portal dostupan je na internetskoj stranici http://hjp.srce.hr.

periodu između početka 1994. i kraja 1995. godine (Ide i Véronis, 1994). U ostvarenju projekt sudjelovao je konzorcij sveučilišta, instituta i komercijalnih tvrtki. Jedan od rezultata bila je i definicija norme za zapis i označavanje cijelih korpusa, dokumenata unutar korpusa, tekstova, poglavlja, odlomaka, rečenica i, onog što je ovom radu najzanimljivije, norme za označavanje vrsta riječi i njihovih morfoloških svojstava. Izvorna norma definirala je pravila za označavanje šest jezika: engleskog, njemačkog, francuskog, španjolskog, talijanskog i nizozemskog.

Nešto kasnije, 1998. godine, norma je proširena definicijom MULTEXT-Easta, grane projekta u kojem su dodatno definirana pravila i skup oznaka za označavanje riječi u jezicima tzv. Istočne Europe. Tada je uključeno šest dodatnih jezika: bugarski, češki, estonski, mađarski, rumunjski i slovenski (Dimitrova et al., 1998). Hrvatski jezik je dodan je u normu 2001. godine. Norma MULTEXT posebna je jer je skup oznaka dizajniran tako da se oznake maksimalno, koliko se to može postići u jezicima s vrlo različitim gramatikama, poklapaju, tvore harmoničan skup kako bi ih se kasnije što lakše moglo koristiti u analizi paralelnih korpusa.

Ulazak u iscrpan opis norme ovdje je preskočen, no neke osnovne informacije o inačici norme za hrvatski jezik ipak je korisno navesti. Norma za morfosintaktičke oznake MULTEXT-East u svojoj prilagodbi za hrvatski jezik svaku riječ opisuje jednom složenom oznakom. Ta se oznaka sastoji od osnovne podoznake, koja grubo odjeljuje vrste riječi, i niza vrijednosti pojedinih atributa te riječi. Za hrvatski se koristi dvanaest osnovnih podoznaka:

1. N za imenice (engl. *noun*),
2. V za glagole (engl. *verb*),
3. A za pridjeve (engl. *adjective*),
4. P za zamjenice (engl. *pronoun*),
5. R za priloge (engl. *adverb*),
6. S za prijedloge (engl. *preposition*, odnosno *adposition*),
7. C za veznike (engl. *conjunction*),
8. M za brojeve (engl. *numeral*),
9. I za usklike (engl. *interjection*),
10. X za ostatke (engl. *residual*),

Vrsta	Atribut	Vrijednosti
Imenica	tip	opća ili vlastita
	rod	ženski, muški, srednji
	broj	jednina, množina
	padež	nominativ... instrumental
	životnost	živa, neživa
Glagol	tip	glavni, pomoćni, modalni...
	oblik	infinitiv, particip...
	vrijeme	futur, perfekt, imperfekt...
	lice	1., 2. ili 3.
	broj	jednina, množina
	rod	ženski, muški, srednji
	stanje	radno ili trpno
Zamjenica	tip	osobna, upitna, posvojna...
	lice	
	rod	
Pridjev	tip	kvalitativni ili posesivni
	stupanj	pozitiv, komparativ...
	lice	
	rod	

Tablica 3.1: Primjeri atributa predviđeni normom MULTEXT-East (za neke vrste riječi)

11. Y za skraćenice (engl. *abbreviation*),

12. Q za čestice (engl. *particle*),

Kao što se to može primijetiti u tablici 3.1 neke vrste riječi imaju bogatiji sustav podoznaka od drugih. Glagoli imaju najveći skup atributa i njihovih vrijednosti, zamjenice su također vrlo razrađene, imenice i pridjevi imaju manji broj atributa i vrijednosti, dok manje podoznaka imaju, primjerice, prijedlozi i skraćenice. Proučavanjem tablice 3.1 i veličina skupova atributa i njihovih vrijednosti može se steći okvirna predodžba o vremenskim zahtjevima označavanja tek jedne riječi u tekstu kao i o zahtjevima na lingvističku spremu ljudskog označivača teksta. Također, može se primijetiti i to da se neke vrijednosti pojedinih atributa mogu pojaviti samo pod uvjetom da su neki drugi atributi već poprimili određene vrijednosti. Drugim riječima, nisu moguće sve kombi-

nacije vrijednosti atributa.

3.2.2. Skup oznaka razvijen u okviru diplomskog rada

Zbog vremenskih ograničenja opisanih u prošlom potpoglavlju nije korišten skup oznaka (engl. *tagset*) koji bi sadržavao sve moguće MSD-e propisane MULTEXT-Eastom. Sustav oznaka razrađen u okviru rada koji je korišten umjesto punih MSD-ova sadrži manji broj oznaka koje, naravno, nose znatno manje informacije o riječima koje opisuju. Skup sadrži 11 od ukupno 12 temeljnih oznaka vrsta riječi kakve se pojavljuju normi MULTEXT-East. Oznaka koje nedostaje je oznaka za ostatak riječi (engl. *residual*). U hrvatskom jeziku ostaci mogu biti, primjerice, nastavci koji se dodaju na složenu skraćenicu poput nastavka *-ovaca* u riječi *FSB-ovaca*. U postupku rastavljanja rečenica na značke ostvarenom u knjižnici koda *MLTools*[4], korištenoj u programskom ostvarenju ovog rada, nastavak dodan na složenu skraćenicu poput *FSB* ne odvaja se od skraćenice jer on nosi bitan dio informacije o njoj – padež, broj ili pak informaciju o tome radi li se ustvari o posvojnom pridjevu poput *FSB-ovčev*. S druge strane, korišteni skup oznaka sadrži i velik broj oznaka kakve nisu predviđene normom MULTEXT-East, a nose važne informacije. Oznake koje su izostavljene u MULTEXT-Eastu, a sadržane u skupu oznaka korištenom u ovome radu su oznake:

- za simbole (npr. simbol za euro €),
- za tzv. *ne-riječi* poput kemijskih i sličnih formula,
- za internetske poveznice (npr. *https://mail.google.com/mail/*),
- za označavanje početka i kraja citata, tj. upravnog govora,
- za riječi iz stranih jezika,
- oznake kojima se razlikuju jednostavne (npr. *prof.*) od složenih skraćenica (npr. *NATO*) i
- oznake kojima se razlikuju različite funkcije interpunkcijskih znakova poput točke

Osim toga ovaj skup dodatno razrađuje oznake za određene vrste riječi. Tako se razlikuju opće od vlastitih imenica, glavni od rednih brojeva te pomoćni glagoli od glavnih. Također se posebnim oznakama označuju glagoli u infinitivu, imperativu, glagolskom pridjevu radnom i trpnom, kao i glagolskom prilogu sadašnjem i prošlom.

[4]MLTools je programska implementacija brojnih alata potrebnih za računalnu obradu prirodnog jezika, poput spomenutog alata za segmentaciju teksta na rečenice i značke, razvijena u Laboratoriju za tehnologije znanja (KTLab) na Fakultetu elektrotehnike i računarstva u Zagrebu.

Skup oznaka OZNAKE1

Prvotni skup oznaka, prikazan u dodatku A, odabran je tako da se njime mogu ručno razriješiti samo višeznačnosti riječi (višeznačnost u smislu: riječ je moguće označiti s više od jedne oznake) koje je teško razriješiti automatski. U nastavku će se za ovaj skup koristiti ime *OZNAKE1*. Neke je riječi, čiju cjelovitu paradigmu[5] nije teško samostalno zapisati, lako strojno detektirati. Primjer su pomoćni glagoli biti i htjeti. Stoga prvotni skup oznaka ne predviđa posebne oznake za, recimo, pomoćne glagole. Umjesto toga, njih se označava nekom od sedam oznaka za glagole, a naknadno se automatiziranim postupkom utvrđuje radi li o pomoćnom glagolu. Skup OZNAKE1 definiran je i opisan u tablici u dodatku A. Ondje je uz svaku oznaku dan i primjer riječi ili druge značke koja se tom oznakom označava. Neka od tih objašnjenja oznaka i primjera preuzeti su iz (Barić et al., 2005).

Pokusima opširno opisanim u poglavlju 6.3, pokazalo se da u korpusu na kojem su provedeni nije bilo pojava internetskih poveznica pa se tako nijednom nije pojavila oznaka URI. Osim toga, posebni tip zamjenica – upitna zamjenica – bila je izdvojena zbog svoje značenjske važnosti. Upitne se zamjenice, naime, pojavljuju samo u upitnim rečenicama. No u rečenicama novinskih tekstova upitne su rečenice vrlo, vrlo rijetke pa se u označenom korpusu nije pojavila niti jedna upitna zamjenica. Jedna pojava u hrvatskom jeziku ponukala je na uvođenje posebne oznake. U primjeru „*jedna od industrijski razvijenih zemalja*“ riječ *razvijenih* smatra se pridjevom iako formalno ima oblik glagolskog pridjeva trpnog. Prema tome bi se riječ *industrijski* trebala smatrati također pridjevom, budući da se prilozi u pravilu susreću uz glagole. Ovakva pojava pridjeva koji nalikuje prilogu u prvom je skupu oznaka dobila posebnu oznaku DL. Redni brojevi prikazani slovima označavani su oznakom BR, dok su redni brojevi prikazani brojkom i točkom označavani nizom oznaka BG (glavni broj) i PTR (točka poslije rednog broja).

Skup oznaka OZNAKE2

U drugoj iteraciji razrade skupa oznaka, opisanoj dalje u tekstu, oznaka DL je odbačena, tj. slijepljena s oznakom za priloge. Nakon neuspješne potrage za objašnjenjem te pojave u literaturi koja se bavi hrvatskim jezikom, odabrana je interpretacija koja riječi poput ranije opisanog *industrijski* proziva prilozima. Znakovi kojima se u tekstu, često nedosljedno, uvode upravni govor i citati (znakovi " ’ „“« ») isti su kao kao

[5] *Paradigma* neke riječi u ovom kontekstu označava sve oblike te riječi. npr. za pridjeve su to svi oblici dobiveni variranjem roda, broja, padeža, stupnja i definitnosti.

znakovi kojima se zapisuje da se određena riječ treba uzeti u prenesenom značenju (npr. *„ozbiljan student“*). Oznake PCP, PCK, PNO i PNZ kojima su se označavale sve ove varijante navodnika, slijepljene su u jednu oznaku – PNAVOD. Motivacija za ovaj potez bila je činjenica da HMM, stohastički model u jezgri ostvarenog strojnog označivača, ne uzima u obzir riječi koje se pojavljuju desno od promatrane ili dalje od dva mjesta ulijevo. Zbog tog ograničenja on ne može razaznati radi li se o navodniku u službi izricanja korištenja prenesenog značenja ili citata. Također ne može razabrati otvarajući znak navođenja od zatvarajućeg, osim u slučajevima kada se oni nalaze na početku ili na kraju rečenice. Istom logikom sljepljene su oznake PZO i PZN u jednu jedinu. Oznaka PZO služila je za odvajajuće zareze, koji odvajaju, recimo, surečenice, a oznaka PZN služila je za nabrajajuće zareze, zareze korištene pri redanju dijelova rečenice koji nisu ni umetnute riječi ni pravilne surečenice, već često kratke imenske fraze. Na samom kraju ovog rada nalazi se i dodatak B u kojem se mogu vidjeti rezultati druge iteracije razrade skupa oznaka.

Skup oznaka OZNAKE-RED

Istraživači koji dolaze iz područja računalne znanosti i koji razvijaju strojne sustave za obradu prirodnog jezika, pa tako i one za označavanje vrsta riječi, u pravilu se oslanjaju na jezične resurse koji su ranije bili prikupljeni i označeni. Ovaj posao obično obavljaju lingvisti ili drugi ljudski označivači istrenirani za takav zadatak. Kako to nije bio slučaj u ovom radu, a susret sa zadatkom ručnog označavanja teksta pokazao se vremenski zahtjevnim zato što je nalagao često i dugotrajno referiranje na dostupnu literaturu (gramatike i pravopisi hrvatskog jezika) u kojoj se povremeno nedostajali opisi nekih pojava u jeziku, u obzir je uzeta i ideja pojednostavljenja skupa oznaka za neke vrste riječi iz zatvorenog skupa riječi. Bitno je naglasiti da su u pokusima provedenim na reduciranom skupu oznaka neke riječi, iako u različitim kontekstima pripadaju različitim vrstama, uvijek, prema uvedenoj konvenciji, označivane istom oznakom. Riječ *blizu*, na primjer, može biti i prilog i prijedlog. U rečenici poput *"Stanujem blizu."*, radi se o prilogu, dok se, ako poslije te riječi slijedi imenica u genitivu, kao u *"To se dogodilo blizu kuće"*, radi o prijedlogu (HJP). Prema uvedenoj konvenciji neke su riječi, dakle, uvijek označavane istom oznakom. Popis takvih riječi je napravljen, pa se poznavanjem padeža (u ovom se radu nisu označavali padeži riječi) lako može uvesti distinkcija *blizu* kao priloga i *blizu* kao prijedloga. U trećem dodatku, dodatku C, mogu se vidjeti uvedena pojednostavljenja, tj. konvencije o označavanju pojedinih riječi.

3.3. Poteškoće u označavanju i prijedlozi za daljnju razradu skupa oznaka

Tijekom ručnog označavanja pronađeni su razni problematični slučajevi koji su, iako su, grubo govoreći, bili predviđeni definiranim skupom oznaka, poticali na naknadno uvođenje dodatnih oznaka. Radilo se o značkama čijim su se dodijeljenim oznakama mogle izreći primjedbe. Često za takve riječi vrijede dvostruka pravila. Drugim riječima, i jedna i druga oznaka neke riječi smatra se točnom, sudeći po informacijama Hrvatskog jezičnog portala. U nastavku se može pročitati neke upadljivije nedosljednosti.

1. Riječi poput *pomoću* označavaju se kao prijedlozi a zapravo se radi o imenici koja se u tom padežu (ovdje instrumental) ustalila u jeziku.

2. Kombiniranje povratnog i prijelaznog oblika glagola kao u rečenici „*Kad ga se pogleda.*“ Glagol *pogledati* u ovoj situaciji nije povratni a nalazi se u konstrukciji koja sadrži povratni zamjenicu *se*. U označavanju punim morfosintaktičkim oznakama ova bi situacija izazvala dodatne probleme.

3. Pridjevi u službi prefiksa poput *društveno* u primjeru *društveno-humanističke znanosti* možda zahtijevaju poseban tretman. U ovom su radu tretirani kao obični pridjevi.

4. Dopušteno je izostavljanje pomoćnih glagola u konstrukcijama s povratnim glagolima. Na primjer, „*ja sam se sklonio*“, „*ti si se sklonio*“ali „*on se sklonio*“. U označavanju vrsta riječi ovo nije važno, ali na to treba obratiti pažnju pri obradi rečenice na višoj razini, razini glagolskih i drugih fraza.

5. Dopušteno je mijenjati određenost različitih pridjeva u nizu koji opisuju istu imenicu kao u primjeru „*... izazivaju priličan međunarodni interes.*“ Ova pojava može djelovati zbunjujuće na sustav koji bi grupirao sve pridjeve koji opisuju neku imenicu.

6. Poimeničeni pridjevi možda bi trebali imati posebnu oznaku, posebno u slučaju da nije moguće utvrditi da se radi o poimeničenom pridjevu provjerom postoji li u bliska imenica koju bi on opisivao. Primjer: „*Razglednice šalju i stari i mladi.*“ Srećom, poimeničeni pridjevi se u pravilu pojavljuju u određenom obliku.

7. Rečenice poput „*Dobro je premazati to nečim tamnim.*“ sadrže pridjev bez imenice, u ovom slučaju pridjev *dobro*, koji označava neku neizrečenu stvar. Grubljom analizom moglo bi se zaključiti i da se radi o prilogu budući da konstrukcija prati oblik u kakvom se pojavljuju i prilozi – pridjev u srednjem rodu u nominativu uz glagol.

8. U tekstovima, posebno novinskim, kao veznici se često koriste riječi kojima je primarna funkcija nije veznička (*ni, samo, kako, kao, kad, poslije*). Najčešće se ova uporaba može vidjeti u rečenicama koje sadrže neupravni govor ili neki oblik parafraziranja.

9. Ponekad se koristi srednji rod koji ne odgovara imenici u množini nego broju ispred nje: „*Uhićeno je deset osoba.*“

10. U novinskim tekstovima često se može sresti krnji glagolski oblik, na primjer krnji perfekt koji sadrži samo glagolski pridjev radni bez pomoćnog glagola. Primjerice, u rečenici: „*Sudionici pozvali na cjelovite ustavne promjene.*“

4. Označivač zasnovan na skrivenom Markovljevom modelu

Za ovaj je rad odabran pristup zasnovan na skrivenom Markovljevom modelu, tj. HMM-u, zato što su za njega već obavljena neka istraživanja – (Goldwater i Griffiths, 2007; Goldberg et al., 2008), koja su koristila nenadzirano učenje. Nadogradnja iskustva stečenog u tim radovima bila je primamljiva zbog nepostojanja dostupnog označenog korpusa na hrvatskom jeziku. Ipak je, budući da je u okviru rada ipak obavljeno ručno označavanje jednom manjeg dijela korpusa, implementiran HMM koji uči *nadzirano*.

4.1. Skriven Markovljev model

Objašnjavanje skrivenog Markovljevog modela dobro je započeti s s objašnjenjem općenitog Markovljevog procesa. Općeniti Markovljev proces diskretan je slučajni proces koji zadovoljava Markovljevo svojstvo (Manning et al., 1999). Za Markovljevo svojstvo ponekad se koristi i izraz Markovljeva pretpostavka. Markovljevo svojstvo naziva se još i bezmemorijskim svojstvom stohastičkog procesa i može se izreći na sljedeći način:

> „Stohastički proces zadovoljava Markovljevo svojstvo ako razdioba uvjetne vjerojatnosti budućih stanja procesa, uz poznata prošla stanja i trenutno stanje, ovisi samo o trenutnom stanju.“

Drugim riječima, Markovljevo svojstvo kaže da nije važno na koji je način sustav dosegao trenutno stanje. U praksi se ova pretpostavka neznatno modificira tako što se uvodi ovisnost trenutnog stanja o jednom ili dva protekla stanja. Na trenutno stanje utječu samo jedno ili dva prošla stanja, često nazivana *ograničenim horizontom* (engl. *limited horizon*) a sva ranija stanja se ignoriraju.

Izmjena trenutnog stanja u modelu može se predočiti i kao slučajna varijabla koja, kako vrijeme prolazi, poprima različite slučajne vrijednosti. Dakle, ti se prijelazi do-

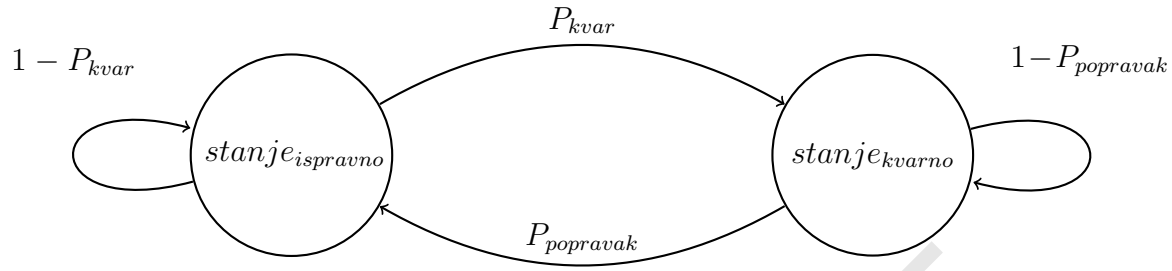
	Stanja sustava vidljiva	Stanja sustava djelomično vidljiva
Sustav je autonoman	Markovljev lanac	Skriven Markovljev model
Sustav je kontroliran	Markovljev proces odlučivanja	Djelomično vidljivi Markovljev proces odlučivanja

Tablica 4.1: Vrste Markovljevih modela

gađaju autonomno, sustav radi sam. Ovaj se detalj spominje zato što postoji i vrsta Markovljevih modela koji modeliraju kontrolirane sustave, sustave u kojima u svakom stanju donositelj odluke (više ili manje inteligentan stroj ili pak čovjek) odlučuje o akciji koja će se primijeniti za prijelaz u sljedeće stanje.

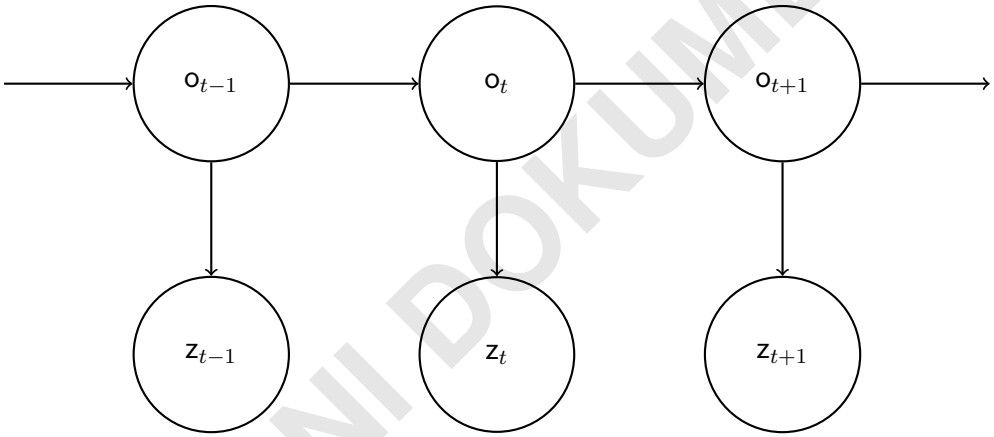
U praksi se često može susresti korištenje različitih naziva (Markovljev model, Markovljev proces, Markovljev lanac) za opisivanje općenitog Markovljeva modela. Markovljev proces i Markovljev model smiju se koristiti kako sinonimi, dok termin *Markovljev lanac* označava posebnu vrstu Markovljeva modela (Manning et al., 1999). Preciznija kategorizacija vrsta Markovljevih modela prikazana je u tablici 4.1. Ako su stanja kroz koja opisani autonomni sustav prolazi vidljiva, tada se govori o Markovljevom lancu, a ako stanja sustava nisu vidljiva nego se mogu samo registrirati nekakva opažanja kakva pojedina stanja emitiraju, tada se radi o skrivenom Markovljevom modelu. Primjer Markovljeva lanca je model ponašanja sustava s komponentama koje se kvare i popravljaju tijekom radnog vijeka sustava. Takav lanac može se vidjeti na slici 4.1. Stanja su ovdje sva vidljiva, jer je prijelaz u kvarno stanje uvijek očit promatračima, kao i povratak u ispravno stanje izvođenjem popravka. Sustav se u danom trenutku nalazi u nekom od stanja. Prijelazi iz stanja u stanje nisu uvijek deterministički definirani. Iz trenutnog se stanja može prijeći u jedno od nekoliko stanja u koja postoje prijelazi iz trenutnog. Uz to, neki su prijelazi vjerojatniji, pa će se shodno tome i češće događati.

U skrivenom Markovljevu modelu poznata su stanja koja sustav u teoriji može poprimiti, no za rada sustava nije moguće vidjeti slijed stanja kroz koja sustav stvarno tada prolazi. Umjesto toga, moguće je primijetiti tek neke pojave koje pojedina stanja emitiraju. Na slici 4.2 sustav prolazi kroz stanja o_{t-1}, o_t i o_{t+1} emitirajući pritom redom opažanja z_{t-1}, z_t i z_{t+1}. Opažanje koje emitira neko stanje također nije deterministički određeno. U nekom je stanju moguće izazvati jednu od više emisija. Neke su emisije



Slika 4.1: Primjer Markovljevog lanca

.



Slika 4.2: Prikaz stanja i emisija skrivenog Markovljevog modela

.

vjerojatnije od drugih, pa se u dužem izvođenju sustava one češće pojavljuju.

4.1.1. Formalan opis skrivenog Markovljevog modela

Raniji tekstovni opis svojstava skrivenog Markovljeva modela može se koncizno zapisati matematičkom notacijom. Formalno, skriveni Markovljev model prvog reda potpuno se definira trojkom (Z, O, μ). Pri tome su elementi trojke (Adler, 2007):

$Z = \{z_1, z_2, ...z_{N_Z}\}$... skup znački –riječi i znakova iz jezika (opažanja HMM-a),

$O = \{o_1, o_2, ...o_{N_O}\}$... skup oznaka (stanja HMM-a),

$\mu = (\Pi, P, E)$... vjerojatnosni model.

N_Z ovdje označava ukupni broj poznatih znački. Pod izrazom *poznate značke* ovdje se podrazumijeva sve moguće riječi i druge pojavnice koje su se pojavile u dijelu korpusa

korištenom za učenje. Ovaj skup poznatih riječi nekad se naziva i *leksikonom*. Uz to, oznaka N_O označava broj svih mogućih oznaka, drugim riječima, N_O označava kardinalnost skupa oznaka definiranih u poglavlju 3.2.2.

Vjerojatnosni model μ opisuje vjerojatnosti prijelaza iz stanja u stanje, tj. iz oznake u oznaku, i emisija znački jezika kao što se to vidi u sljedećim definicijama:

$$\Pi = \{\pi_i, 1 \leqslant i \leqslant N_O\} \text{ ... vjerojatnost pojave oznake } o_i \text{ na početku rečenice} \tag{4.1}$$

$$P = \{p_{ij}, 1 \leqslant i, j \leqslant N_O\} \text{ ... vjerojatnost prijelaza iz oznake } o_i \text{ u } o_j \tag{4.2}$$

$$E = \{e_{i,z_l}, 1 \leqslant i \leqslant N_O, 1 \leqslant l \leqslant N_Z\} \text{ ... vjerojatnost emisije značke } z_l \text{ iz oznake } o_i. \tag{4.3}$$

Ako se uvedu neke dodatne oznake, može se pobliže definirati vjerojatnosti prijelaza poput p_{ij} u skrivenom Markovljevom modelu prvog reda i p_{ijk} u HMM-u drugog reda. Neka T označava ukupni broj znački u rečenici, neka z_t označava značku na mjestu t u rečenici i neka o_t označava oznaku ručno pridijeljeno znački na t-tom mjestu u rečenici. Tada se vjerojatnosti mogu lako formalno definirati. Vjerojatnost p_{ij} zapravo označava

$$p_{ij} = P(\mathsf{o}_t = o_j | \mathsf{o}_{t-1} = o_i),$$

dok vjerojatnost e_{i,z_l} označava

$$e_{i,z_l} = P(\mathsf{z}_t = z_l | \mathsf{o}_t = o_i).$$

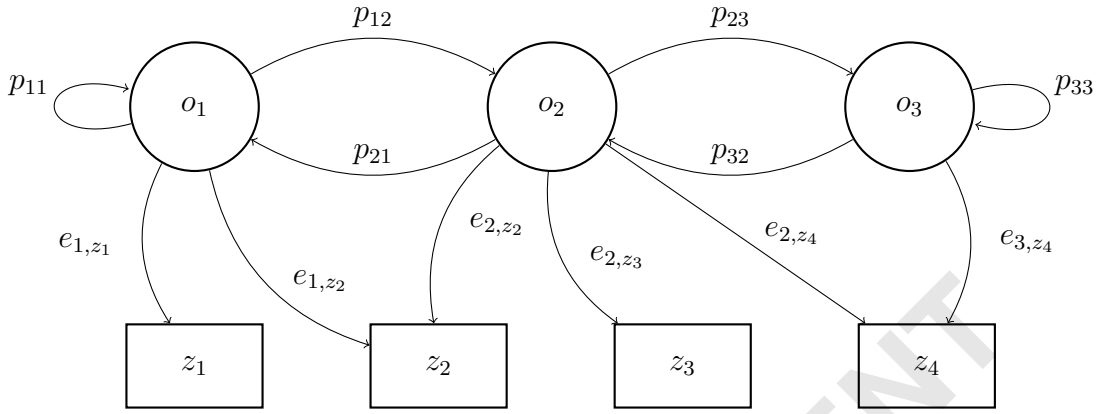
Ako se pak koristi HMM drugog reda, tada se u obzir uzimaju i oznake na prethodnom mjestu, pa se tako definiraju vjerojatnosti p_{ijk} i e_{ij,z_l}:

$$p_{ijk} = P(\mathsf{o}_t = o_k | \mathsf{o}_{t-1} = o_j, \mathsf{o}_{t-2} = o_i)$$

$$e_{ij,z_l} = P(\mathsf{z}_t = z_l | \mathsf{o}_t = o_j, \mathsf{o}_{t-1} = o_i).$$

Definicije (4.1) govore o skrivenom Markovljevom modelu prvog reda. U označavanju vrsta riječi svi noviji radovi koriste model drugog reda, pa se tako on koristi i u ovom radu. Označivače koji koriste HMM prvog reda naziva se bigramskim označivačima, dok se one koji koriste HMM drugog reda naziva trigramskim označivačima.

Slika 4.3 prikazuje jednostavan skriveni Markovljev model. Model se sastoji od tri stanja o_1, o_2 i o_3. Na strelicama su prikazane vjerojatnosti prijelaza iz stanja u stanje i vjerojatnosti emisije pojedinih znakova. Iz svakog se stanja emitira jedan znak, pa se tako iz stanja o_2 može emitirati jedan od tri moguća znaka (z_2, z_3 ili z_4), dok se iz stanja o_3 uvijek emitira znak z_4.



Slika 4.3: Primjer skrivenog Markovljevog modela prvog reda

.

4.1.2. Neke primjene skrivenog Markovljeva modela

Skriveni Markovljevi modeli primjenjuju se na različite probleme. U raspoznavanju govora problem je iz zvučnog zapisa, grafa s vremenom na osi x i intenzitetom zvuka na osi y, koji prikazuje snimljeni zvuk, raspoznati koji su stvarni slogovi teksta izgovoreni. Uz dani zvučni zapis potrebno je razlučiti koje je, možda vrlo slične, slogove govornik stvarno izgovorio. Je li, na primjer, izgovorena riječ *bas* ili *pas*? U ovakvoj primjeni sustav promjenom vremena prolazi kroz stanja i pritom emitira opažanja. Stanja su ovdje stvarni izrečeni slogovi i oni su nepoznati. Poznata su samo vidljiva opažanja, u ovom slučaju zvučni zapis izgovora tih slogova. Bioinformatika još je jedno područje u kojem se HMM-i često koriste. Jedna primjena HMM-a u bioinformatici je problem pronalaska gena. Cilj je pronaći kodirajuća i nekodirajuća područja niza nukleotida poput adenina, citozina, guanina i timina u deoksiribonukleinskoj kiselini (DNK). Ovdje domena nije vrijeme, nego pozicija u nizu nukletida u DNK. Dakle, na svakoj poziciji u nizu nalazi se jedan nukleotid čije stanje poprima vrijednost iz skupa { *„nukleotid je dio nekodirajućeg područja“*, *„nukleotid je dio direktno kodirajućeg niza u DNK“*, *„nukleotid kodira protein“* } (Lukashin i Borodovsky, 1998). Ova su stanja nepoznata, no opažanja, konkretne pojave adenina, citozina, guanina i timina, vidljiva su opažaču.

4.2. Princip rada označivača zasnovanog na skrivenom Markovljevom modelu

Napokon se može opisati primjena skrivenog Markovljevog modela na problem označavanja vrsta riječi. Stanja su u ovoj primjeni vrste riječi predstavljene njihovim oznakama uvedenim u poglavlju 3.2.2. Stanja opažaču nisu poznata. Umjesto njih, opažač može vidjeti sâme riječi. Sustav u stvarnosti prolazi kroz stanja kao u sljedećem primjeru:

IO IO J IO GPOM GIPOM J BG IO PKD.

Ove oznake vrsta riječi nisu poznate. Programirani automatizirani označivač treba ih sam otkriti tako da pronađe slijed oznaka koji je najvjerojatniji uz dana opažanja, tj. značke rečenice. Dakle, označivaču su poznate samo riječi koje svako stanje emitira:[1]

Cijena benzina u ponoć će skočiti preko 10 kuna?

Označavanje se obavlja na razini rečenice. Za svaku se rečenicu posebno odabire najvjerojatniji niz oznaka i taj se postupak ponavlja za cijele odlomke, dokumente i, konačno, cijeli korpus. Najvjerojatniji se niz oznaka pronalazi pomoću Viterbijevog algoritma, detaljno opisanog u poglavlju 5.1. Kako bi se algoritam mogao izvesti, potrebno je imati poznate sve vjerojatnosti prijelaza iz stanje u stanje, tj. iz oznake u oznaku, kao i vjerojatnosti da pojedina oznaka emitira neko opažanje, tj. značku.

4.2.1. Procjena vjerojatnosti u skrivenom Markovljevom modelu

Postupak pronalaska, za rad skrivenog Markovljeva modela neophodnih, vjerojatnosti prijelaza i emisija, koje se negdje naziva i *parametrima* HMM-a (Goldberg et al., 2008), može se smatrati svojevrsnim učenjem modela. Tako naučen model, odnosno označivač, može se primijeniti na označavanje riječi u novim, neviđenim rečenicama. Označivač vrsta riječi, kao i tipičan klasifikator, prvo prolazi fazu učenja koje se obavlja na većem skupu označenih podataka i, kasnije, fazu testiranja, tj. označavanja koja se obavlja na drugom, manjem skupu označenih podataka.

Nadgledano učenje vjerojatnosti

Nadgledano učenje vjerojatnosti prijelaza i emisija obavlja se vrlo jednostavno. Učenje se svodi na procjenu vjerojatnosti na temelju broja pojava prijelaza iz oznake u oznaku

[1] Preuzeto s www.danas.hr 7.3.2011.

i pojava emisija određenih znački iz određenih oznaka. U literaturi se za ovaj postupak najčešće koristi izraz *procjena najveće izglednosti* (engl. *maximum likelihood estimation* – MLE). Ovaj složeni naziv ne odražava dobro izrazitu jednostavnost postupka. Vjerojatnost da prva oznaka u rečenici bude o_i – π_i dobiva se na sljedeći način:

$$\pi_i = \frac{B_{i,poč}}{B_{reč}} \tag{4.4}$$

Pri čemu je $B_{i,poč}$ broj rečenica u kojima je na prvom mjestu oznaka o_i, a $B_{reč}$ ukupni broj rečenica. Vjerojatnost prijelaza iz oznake o_i u oznaku o_j, tj. vjerojatnost p_{ij}, dobiva se na sljedeći način:

$$p_{ij} = \frac{B_{ij}}{B_i}. \tag{4.5}$$

Pri tome je B_{ij} ukupni broj pojava oznake o_j nakon oznake o_i. Drugim riječima, radi se o ukupnom broju pojava bigrama (o_i, o_j) u cijelom korpusu. B_i ovdje označava ukupni broj pojava oznake o_i u cijelom korpusu. Ekvivalentna formula za HMM-e drugog reda, gdje se promatra vjerojatnost prijelaza u oznaku o_k nakon što je sustav redom prošao kroz oznake o_i i o_j, je:

$$p_{ijk} = \frac{B_{ijk}}{B_{ij}}, \tag{4.6}$$

Gdje je B_{ijk} broj pojava oznake o_k nakon bigrama (o_i, o_j), a B_{ij} broj pojava tog bigrama. Vjerojatnosti emisija dobivaju se formulom

$$e_{i,z_l} = \frac{B_{i,z_l}}{B_i}, \tag{4.7}$$

odnosno formulom

$$e_{ij,z_l} = \frac{B_{ij,z_l}}{B_{ij}} \tag{4.8}$$

za slučaj HMM-a drugog reda. Pri tome je B_i broj pojava (unigrama) oznake o_i, B_{ij} broj pojava bigrama (o_i, o_j), a B_{i,z_l} i B_{ij,z_l} su redom broj pojava emisije z_l iz stanja o_i, odnosno broj pojave emisije z_l iz stanja o_j nakon što je sustav u njega došao iz stanja o_i.

Velik broj ovih vjerojatnosti, pogotovo u vjerojatnostima emisija i pogotovo ako se radi o skrivenom Markovljevu modelu drugog reda, jednak je nuli. Ova se pojava naziva rijetkošću podatka (engl. *data sparseness*). Budući da se u računanju najvjerojatnijih slijedova oznaka u rečenici učestalo koristi množenje vjerojatnosti, česta pojava vjerojatnosti jednake nuli uzrokuje pogoršanje uspješnosti označivača. Ovaj fenomen objašnjen je podrobnije u idućem potpoglavlju koje je posvećeno problemu rijetkih podataka i zaglađivanju vjerojatnosti.

4.2.2. Problem rijetkih podataka i zaglađivanje vjerojatnosti

Kratko rečeno, pojava velikog broja nula u matricama koje sadrže vjerojatnosti prijelaza i emisija uzrokuje smanjenje uspješnosti označivača. Ovo se događa zato što u tom slučaju velik broj mogućih kombinacija oznaka znački u rečenici ima jednaku vjerojatnost – vjerojatnost jednaku nuli. U stvarnosti su neke od tih kombinacija oznaka vjerojatnije od drugih, no te se nijanse gube zbog ograničenog korpusa na kojem se provodi učenje označivača. U idealnom slučaju bi korpus bio vrlo velik i prekrivao sve moguće rasporede pojave unigrama i bigrama oznaka znački u tekstu, kao i emisija samih znački iz tih oznaka. No čak i u tom slučaju bi rijetkost podataka bila nezanemariva. Uz skup oznaka koji broji 41 element, i broj poznatih riječi od oko 9000,[2] teorijski je moguće oko 15 milijuna kombinacija dvije oznake i emisije značke iz druge po redu oznake. Naravno, veliki broj teorijskih kombinacija oznaka i emisija u praksi se ne pojavljuju.

Potrebno je, dakle, smanjiti udio elemenata u matricama vjerojatnosti jednakih nuli. To se obavlja zaglađivanjem vjerojatnosti. Grubo rečeno, postupak se može objasniti kao zamjena nekih vrijednosti vjerojatnosti jednakih nuli vrlo malenim vjerojatnostima većim od nule. Naravno, te se vjerojatnosti ne odabiru nasumično nego se u slučaju pojave vrijednosti nula za neki trigram, koriste podaci nižeg reda – vjerojatnosti bigrama i unigrama. Kako je to predloženo u (Thede i Harper, 1999), zaglađene vjerojatnosti prijelaza se računaju prema formuli:

$$p_{ijk} = k_3 \cdot \frac{B_{ijk}}{B_{ij}} + (1 - k_3)k_2 \cdot \frac{B_{jk}}{B_j} + (1 - k_3)(1 - k_2) \cdot \frac{B_k}{N_O}. \tag{4.9}$$

Brojevi B_{ijk}, B_{jk} i drugi u ovoj formuli definiraju se jednako kao u izrazima (4.4) do (4.8). Težinski faktori k_3 i k_2 definirani su kao:

$$k_3 = \frac{\log(B_{ijk} + 1) + 1}{\log(B_{ijk} + 1) + 2} \tag{4.10}$$

i

$$k_2 = \frac{\log(B_{jk} + 1) + 1}{\log(B_{jk} + 1) + 2}. \tag{4.11}$$

Zaglađena vjerojatnost emisije neke značke dana je s

$$e_{ij,z_l} = \frac{\log(B_{ij,z_l} + 1 + 1}{\log(B_{ij,z_l} + 1) + 2} \cdot \frac{B_{ij,z_l}}{B_{ij}} + \frac{1}{\log(B_{ij,z_l} + 1) + 2} \cdot \frac{B_{i,z_l}}{B_i}. \tag{4.12}$$

Kako se vidi u formulama (4.9) i (4.12), ako se neki trigram nikada ne pojavljuje u dijelu korpusa za učenje, tada se vjerojatnost oslanja na pojave bigrama i unigrama

[2] Otprilike ovoliko je riječi zabilježeno pri provođenju učenja na podkorpusu od 15000 znački.

koje taj trigram sadrži. Također, može se vidjeti da i s zaglađivanjem neke vjerojatnosti ostaju jednake nuli, no u dovoljno malenom broju da ne utječu znatno na konačnu uspješnost označivača.

Ključni algoritam za pronalazak niza oznaka za značke neke rečenice, tako da se maksimira vjerojatnost da je on upravo niz ispravnih oznaka, je Viterbijev algoritam. Njegov opis i kratki pseudokod nalaze se u poglavlju 5.1.

4.2.3. Postupanje s nepoznatim riječima

Poseban problem u označavanju vrsta riječi je označavanje nepoznatih znački (engl. *unknown words*), znački na koje se nije naišlo tijekom učenja označivača. Ovakve značke, tzv. značke „izvan leksikona “, ne nalaze se u matricama vjerojatnosti emisija znački. Prema tome, nije moguće koristiti znanje o vjerojatnosti emisije nepoznate značke iz dane oznake. Pristup kojim se za veliki broj nepoznatih riječi i u najnovijim radovima postižu dobri rezultati njihova označavanja je analiza informacija sadržanih u afiksima riječi. Prvi o tome govore (Klein i Simmons, 1963) dok se daljnja razrada ideje može pronaći u (Samuelsson, 1993). U okviru rada opisanog u (Brants, 2000) razvijena je vrlo popularna implementacija HMM-a koji koristi informaciju sadržanu u sufiksima riječi, a čak i najnoviji radovi, npr. (Goldberg et al., 2008; Agić et al., 2009), koriste informacije iz afiksa riječi. U praksi je najpogodnije, pogotovo u hrvatskom jeziku, koristiti samo sufikse kao nositelje informacije, zanemarujući manje korisne prefikse i infikse. Osim sufiksa, kao indikator vrste kojoj neka nepoznata značka pripada, uzimaju se i neke dodatne informacije poput:

- počinje li značka velikim slovom (uz pretpostavku da se ne nalazi na prvom mjestu u rečenici),
- jesu li sva slova u znački velika,
- sadrži li značka crtu

i slično.

Za poznatu značku, označivaču je poznata i vjerojatnost $P(z_l|o_i)$, vjerojatnost da oznaka o_i emitira upravo značku z_l. No ako je neka značka na koju je označivač naišao nepoznata, tada se Viterbijev algoritam oslanja na vjerojatnost $P(\text{značka sadrži sufiks } s_l|o_i)$. Pored matrice E koja sadrži vjerojatnosti emisija znački, uvodi se dodatna matrica N definirana na sljedeći način:

$$N = \{n_{ij,s_l} : n_{ij,s_l} = P(\mathsf{z}_t \text{ sadrži sufiks } s_l|\mathsf{o}_t = o_j, \mathsf{o}_{t-1} = o_i)\}. \tag{4.13}$$

U nadgledanom se učenju, dakle, vjerojatnosti pojave nepoznatih riječi procijenjuju na temelju njihovog nastavka. Treba imati na umu da nastavak riječi, koji se ovdje naziva *sufiksom* nije nužno uvijek lingvistički značajan, ispravan sufiks u punom značenju te riječi (Brants, 2000).

U ovom radu se koriste sufiksi do maksimalne duljine od četiri znaka. Njihove se vjerojatnosti dobivaju analogno onima za poznate riječi. Sve se opet obavlja brojenjem pojava znački s određenim sufiksima kao što se to vidi u

$$n_{ij,s_l} = \frac{B_{ij,s_l}}{B_{ij}}. \tag{4.14}$$

Ovdje B_{ij,s_l} označava broj znački u korpusu za učenje koje koje sadrže sufiks s_l, uz danu oznaku o_j te značke i oznaku o_i prethodne značke. Ovaj se broj dobije tako što se prolaskom kroz sve rečenice u korpusu za učenje svakoj znački pogledaju sufiksi duljine od jedan do četiri znaka i povećaju odgovarajući brojači pojava. U potrazi za najvjerojatnijim slijedom oznaka Viterbijevim algoritmom, vjerojatnost emisije jednostavno se odabire prema pravilu

$$P_{emisije} = \begin{cases} e_{ij,z_l} & \text{ako je riječ poznata} \\ n_{ij,s_l} & \text{ako je riječ nepoznata} \end{cases} \tag{4.15}$$

Matrica N sadrži vjerojatnosti emisije znački koje sadrže neke sufikse, no dodatna svojstva znački mogu se upotrijebiti za postizanje boljih konačnih rezultata označivanja stvaranjem posebnih matrica koje će sadržavati razdiobu vjerojatnosti za posebne podvrste nepoznatih riječi. Tako je moguće izraditi posebnu matricu N_{prvo_veliko} koja bi sadržavala vjerojatnosti emisije značke s velikim prvim slovom (uz zahtjev da se riječ ne nalazi na početku rečenice), matricu N_{sva_velika} koja bi sadržavala razdiobu za značke čija su sva slova velika (npr. akronimi i druge složene skraćenice) i matricu N_{crta} koja bi sadržavala razdiobu za sve značke koje sadrže spojnicu ili crtu.

Rezultati označavanja nepoznatih riječi dodatno su poboljšana uvođenjem naknadne revizije dodijeljenih oznaka. Nakon što označivač zasnovan na HMM-u riječima u skupu za testiranje dodijeli oznake, oznake dodijeljene nepoznatim riječima se provjeravaju. To se obavlja tako što se za svaku nepoznatu riječ pogleda kakvu joj oznaku dodjeljuje lematizacijska komponenta. Ako ta komponenta za riječ predlaže samo jednu oznaku (dakle, ako riječ može biti označena samo jednom jedinom oznakom) tada se, umjesto oznake koju je dodijelio označivač zasnovan na HMM-u, koristi

oznaka koju predlaže lematizacijska komponenta iz MLTools.[3]

[3]Jedan od alata u MLTools je i lematizacijska komponenta čija je funkcija za bilo koji dani oblik riječi pronaći njezin jedan ili više osnovnih oblika, tzv. lema (nominativ jednine za imenice, infinitiv za glagole itd.). Osim toga, lematizacijska komponenta za dani oblik neke riječi može ponuditi sve oznake kojima se on može označiti.

5. Programsko ostvarenje

5.1. Viterbijev algoritam

Najveći dio vremena izvođenja programsko ostvarenje provodi u funkciji koja ostvaruje Viterbijev algoritam pa je stoga on ovdje i prikazan. Cjeloviti opis algoritma i teorijske pozadine na kojoj je on sazdan zahtijevao bi vlastitih nekoliko stranica, pa je ovdje dan samo njegov sažeti matematički zapis.

Ovdje je uputno spomenuti da se na početak svih rečenica, kako bi se sve vrijednosti varijable δ mogle popuniti u HMM-u drugog reda, dodala umjetna početna značka i umjetna početna oznaka vrste. Ove dodane značke i oznake ne nose nikakvu informaciju i ne utječu na stvarne vjerojatnosti. Dodane su samo zato da bi se vjerojatnosti pojave emisije značke z_k iz isključivo trenutne oznake o_i (vjerojatnost $P(\mathsf{z}_t = z_k|\mathsf{o}_t = o_i)$) na prvom mjestu u rečenici uklopio u matricu koja sadržava vjerojatnosti s obzirom na trenutnu i prethodnu oznaku (vjerojatnost $P(\mathsf{z}_t = z_k|\mathsf{o}_{t-1} = o_i, \mathsf{o}_t = o_j)$).

Algoritam se sastoji od dva dijela:

1. indukcije (engl. *induction*) i
2. završetka s iščitavanjem optimalnog puta povratkom kroz put algoritma (engl. *termination and path readout by backtracking*).

U oba se dijela koriste dvije varijable, δ i ψ. Varijabla δ je definirana na sljedeći način:

$$\delta_{ij}(t) = \max_{\mathsf{o}_1,\ldots,\mathsf{o}_{t-2}} (\mathsf{o}_1, \ldots \mathsf{o}_{t-2}, \mathsf{o}_{t-1} = o_i, \mathsf{o}_t = o_j, \mathsf{z}_1, \ldots, \mathsf{z}_t), 2 \leqslant t \leqslant T. \tag{5.1}$$

S $\delta_{ij}(t)$ predstavljen je najbolji, najvjerojatniji niz oznaka počevši od prve oznake o_1 do oznake o_{t-2} koji završava s oznakama o_i i o_j na $(t-1)$-om i t-om mjestu u rečenici. Varijabla ψ se definira kao:

$$\psi_{ij}(t) = \arg \max_{\mathsf{o}_1,\ldots,\mathsf{o}_{t-2}} (\mathsf{o}_1, \ldots \mathsf{o}_{t-2}, \mathsf{o}_{t-1} = o_i, \mathsf{o}_t = o_j, \mathsf{z}_1, \ldots, \mathsf{z}_t), 2 \leqslant t \leqslant T. \tag{5.2}$$

To znači da ψ uvijek sadrži oznaku na poziciji $t-2$ koja prethodi dvjema oznakama na $t-1$ i t.

5.1.1. Indukcija

Indukcija je zapravo postupak popunjavanja svih polja matrice koja predstavlja vrijednosti varijable δ. Zbog dodatne umjetne oznake na početak svih rečenica, brojanje ovdje počinje od 0 umjesto od 1.

Prvo se obavlja inicijalizacija na prvim dvjema pozicijama u rečenici. Na nultoj poziciji postavlja se:

$$\delta_{ij}(0) = d, 1 \leqslant i, j \leqslant N_O \tag{5.3}$$

pri čemu d označava neutralni element s obzirom na operaciju množenja; drugim riječima:

$$d = \begin{cases} 1 & \text{u slučaju množenja vjerojatnosti} \\ 0 & \text{u slučaju zbrajanja logaritama vjerojatnosti.}[1] \end{cases} \tag{5.4}$$

Na prvoj poziciji se obavlja inicijalizacija prema

$$\delta_{ij}(1) = \left\{ \begin{array}{ll} \pi_i e_{ij,\mathsf{z}_1} & \text{ako je } \mathsf{z}_1 \text{ poznata} \\ \pi_i n_{ij,\mathsf{z}_1} & \text{ako je } \mathsf{z}_1 \text{ nepoznata} \end{array} \right\}, 1 \leqslant i, j \leqslant N_O. \tag{5.5}$$

Na svim se sljedećim pozicijama u rečenici, počevši od $t = 2$ do T, indukcija obavlja prema formuli

$$\delta_{jk}(t) = \left\{ \begin{array}{ll} \max\limits_{1 \leqslant i \leqslant N_O} \delta_{ij}(t-1) p_{ijk} e_{jk,\mathsf{z}_t} & \text{ako je } \mathsf{z}_t \text{ poznata} \\ \max\limits_{1 \leqslant i \leqslant N_O} \delta_{ij}(t-1) p_{ijk} n_{jk,\mathsf{z}_t} & \text{ako je } \mathsf{z}_t \text{ nepoznata} \end{array} \right\}, 1 \leqslant j, k \leqslant N_O. \tag{5.6}$$

Pri tome se množenje s faktorom e_{jk,z_t}, odnosno n_{jk,z_t}, treba obaviti izvan petlje koja iterira i.

5.1.2. Završetak i iščitavanje puta

Po završetku indukcije, optimalni niz oznaka iščitava se prolaskom unatrag kroz vrijednosti varijable δ, prateći pritom najveće vjerojatnosti. Postupak je najbolje precizno objasniti matematičkim opisom. Prvo se pronađu dvije završne oznake za koje je postignuta maksimalna vjerojatnost $\hat{P}$ ukupnog niza oznaka za cijelu rečenicu maksimirajući izraz

$$\hat{P} = \max_{1 \leqslant i, j \leqslant N_O} \delta_{ij}(T). \tag{5.7}$$

Te dvije završne oznake se pohranjuju na zadnja dva mjesta niza ô koji će pohranjivati najvjerojatniju kombinaciju oznaka za značke u promatranoj rečenici. Zadnja oznaka

[1] Rad s logaritmima vjerojatnosti objašnjen je u sljedećem poglavlju.

pronađe se kao

$$\hat{\mathsf{o}}_T = \arg_j \max_{1 \leqslant i,j \leqslant N_O} \delta_{ij}(T), \tag{5.8}$$

a predzadnja kao

$$\hat{\mathsf{o}}_{T-1} = \arg_i \max_{1 \leqslant i,j \leqslant N_O} \delta_{ij}(T). \tag{5.9}$$

Nakon toga se čitanjem vrijednosti varijable ψ, idući prema kraju, popunjavaju svi ostali elementi niza $\hat{\mathsf{o}}$

$$\hat{\mathsf{o}}_t = \psi_{\mathsf{o}_{t+1}\mathsf{o}_{t+2}}(t+2). \tag{5.10}$$

Napokon, može se iščitati cjelokupni sadržaj niza $\hat{\mathsf{o}}$, otkrivajući time najvjerojatniji niz oznaka za značke promatrane rečenice.

5.2. Numeričke specifičnosti zadatka

Aritmetički podljev (engl. *arithmetic underflow*) čest je problem u računalnim programima koji provode izračune umnožaka vjerojatnosti. Kao što se to vidi u izrazu (5.6), u Viterbijevom algoritmu provodi se velik broj množenja često vrlo malenih vjerojatnosti. Tipovi podataka s pomičnim zarezom, npr. `float` ili `double` imaju samo ograničenu preciznost i ne mogu pohranjivati pozitivne vrijednosti manje od ε, tzv. strojnog epsilona stroja na kojem se provodi računanje. Sve vrijednosti veće od nule i manje od ε zapisuju se kao nule.

Ako se problem podljeva prikladno ne riješi tada se u matrici koja predstavlja sve vrijednosti varijable δ pojavljuje veliki broj nula, što konačno rezultira time da za veliki broj kombinacija oznaka njihove vjerojatnosti budu međusobno jednake – jednake nuli. Ova pojava efektivno onemogućuje razlikovanje vjerojatnijih od manje vjerojatnih pridjeljivanja oznaka značkama rečenice.

Zbog ovih razloga se umjesto sâmih vjerojatnosti koriste njihovi logaritmi. Pri tome se koristi prednost logaritmiranja umnožaka:

$$\log(a \cdot b) = \log a + \log b. \tag{5.11}$$

Iz (5.11) se vidi da je problem podljeva riješen zamjenom umnožaka malenih vjerojatnosti zbrajanjem njihovih logaritama. U programskom se ostvarenju stoga svaka pojava neke vjerojatnosti, primjerice vjerojatnosti p_{ijk} (prijelaz u oznaku o_k nakon oznaka o_i i o_j), umjesto nje koristi njezin logaritam – $\log p_{ijk}$. U ovoj se implementaciji koristi prirodni logaritam $\log_e x$, a skraćeno ga se zapisuje kao $\log$. Tako množenje postaje

$$p_1 \cdot p_2 \longrightarrow \log p_1 + \log p_2 \tag{5.12}$$

a zbrajanje, prema (Aji i McEliece, 2002), postane

$$p_1 + p_2 \longrightarrow \hat{p} + \log(e^{\log p_1 - \hat{p}} + e^{\log p_2 - \hat{p}}) \text{ gdje je } \hat{p} = \max\{p_1, p_2\}. \tag{5.13}$$

U ovoj implementaciji se nigdje ne obavlja zbrajanje vjerojatnosti, no u algoritmu Bauma i Welcha, kojeg se koristi za nenadgledano učenje parametara HMM-a, kao u (Goldberg et al., 2008), potrebno je moći zbrajati vjerojatnosti, tj. njihove logaritme.

5.3. Komentar programskog ostvarenja

Prvotna vizija ovog diplomskog rada predviđala je korištenje nenadziranog učenja zbog nedostupnosti označenog korpusa. U skladu s takvim planom izrađena je implementacija većeg dijela modela označivača opisanog u (Goldberg et al., 2008). Ručnim označavanjem vlastitog korpusa otvorila se mogućnost korištenja nadgledanog učenja pa je veća količina programskog koda ostala, bar u ovom trenutku, neiskorištena. Daljnji rad mogao bi uključivati razvoj označivača koji uči nenadzirano, poduprt značajnim, ali ipak ne dovoljno velikim označenim korpusom.

6. Vrednovanje uspješnosti

6.1. Korpus

Korpus nad kojim je provedeno učenje skrivenog Markovljeva modela i ocjenjivanje uspješnosti njegovog rada sastojao se od novinskih članaka iz Vjesnika. Vjesnik je politički dnevni list koji izlazi u Zagrebu od 1940. godine. U svom modernom obliku, svako je izdanje Vjesnika sastavljeno od članaka podijeljenih u 12 rubrika označenih s njihovim skraćenim imenima (u zagradama):

1. Crna kronika (crn),
2. Gospodarstvo (gos),
3. Komentari (kom),
4. Kultura (kul),
5. Sport (spo),
6. Sa svih strana (sss),
7. Tribina (sta),
8. Teme (tem),
9. Događaji (unu),
10. Svijet (van),
11. Zagreb i županija (zag) i
12. Život (pis).

Svaka od ovih rubrika sadrži određeni broj članaka broj članaka. Neke rubrike ipak imaju tendenciju sadržavati znatno više članaka od drugih. Tako rubrika „Sport “vrlo

često sadrži i preko 20 članaka, dok rubrika „Komentari “najčešće sadrži tek tri do četiri članka.

Korpus koji je bio na raspolaganju sadrži neobrađena sva dnevna izdanja Vjesnika iz proteklih deset godina, počevši od svibnja 1999. sve do studenog 2009. godine. Za pokuse u ovom radu korišten je ipak samo dio korpusa. Sve su značke u tom dijelu korpusa ručno označene pa će ga se u nastavku kratko nazvati *označenim korpusom*. Dio korpusa koji je ručno označen odabran je na sljedeći način:

- rečenice su odabirane počevši od zadnjeg dostupnog izdanja (1. studenog 2009) prema starijim izdanjima,
- iz svakog je izdanja odabirano po 12 rečenica,
- svaka od tih 12 rečenica iz jednog izdanja bila je odabrana iz jedne od 12 rubrika, tako da ni iz jedne rubrike nije uzeto više od jedne rečenice,
- odabirana je rečenica na nasumičnom mjestu u nasumično odabranom članku unutar rubrike.

Ovakav odabir rečenica rezultira jednakom zastupljenošću rečenica iz svake od rubrika, poništavajući tako relativnu veličinu rubrika poput „Sporta “koji često sadrži velik udio vrlo krnjih rečenica, primjerice kada se u cijeloj rečenici samo nabrajaju sportaši i njihovi rezultati na nekom natjecanju. Osim toga, članci su odabirani nasumično umjesto, recimo, stalnog odabiranja prvih, udarnih članaka u rubrici koje vjerojatno tendenciju imaju pisati istaknutiji autori. Nasumični odabir rečenice iz članka bolji je od konstantnog uzimanja prve ili zadnje rečenice zato što u novinskim tekstovima prve rečenice, a i zadnje često prate okvirno sličnu formu.

Izdvojeni i označeni korpus sadrži ukupno 20000 znački i seže od zadnjeg dostupnog izdanja (izdanje od 1. studenog 2009.) sve do izdanja od 2. rujna 2009.

Okvirna brzina koju može postići ljudski označivač nakon što je proveo nekoliko dana intenzivno se upoznavajući sa sustavom oznaka i lingvističkim specifičnostima vrsta riječi u hrvatskom jeziku je oko 1000 znački u sat vremena. Ovisno o stupnju fokusiranosti označivača, ova brojka može oscilirati do nekoliko stotina znački na sat. Postupak bi se mogao ubrzati modifikacijom sustava za ručno označavanje dodavanjem funkcije koja za sve riječi, pogotovo one iz skupa zatvorenih riječi, ljudskom označivaču daje izbor od svega nekoliko mogućih oznaka za tu riječ.

U tablici 6.1 mogu se podgledati udjeli riječi u tekstu koje imaju jednu jedinu moguću oznaku naspram onih koje je, ovisno o rečeničnom kontekstu, moguće označiti dvjema, trima ili pak četirima različitim oznakama. Ovo je mjerenje provedeno na oko milijun znački (iz 40000 rečenica) iz neoznačenog dijela Vjesnikovog korpusa. Reče-

Broj mogućih oznaka	Broj znački	Udio u korpusu [%]
0	37846	4,34
1	690981	79,22
2	123668	14,18
3	19273	2,10
4	418	0,05

Tablica 6.1: Raspodjela riječi s više mogućih oznaka (prema MLTools)

nice su redom uzimane počevši od najstarijeg dostupnog izdanja. Postupak se sastojao u prolasku kroz sve riječi u tekstu te u korištenju lematizacijske komponente koja ima mogućnost za zadanu riječ, u kojem god ona obliku bila, dati sve oznake kojima bi ta riječ mogla biti označena. Odabir oznake nije nije ovisio o kontekstu dane značke, već je istoj znački uvijek dodjeljivana ista oznaka. Za brojanje oznaka koje u teoriji ista riječ može poprimiti korištena je lematizacijska komponenta sadržana u knjižnici koda MLTools. Ta je lematizacijska komponenta proizvod rada provedenog u (Šnajder, 2010). Maksimalni broj oznaka koje su moguće za neku riječ ovdje je četiri. U ovom pokusu nisu uzimane u obzir sve značke (uključujući interpunkcijske znakove i sl.) nego samo stvarne riječi kakve lematizacijska komponenta može analizirati. Može se primijetiti da se ipak pojavio relativno malen dio riječi (4%) za koje nije ponuđena ni jedna moguća oznaka.

Vrlo sličan pokus proveden je nad označenim korpusom. Ovdje nije korištena lematizacijska komponenta nego su korištene ručno dodijeljene oznake. Pri tome se koristio skup oznaka OZNAKE2. Za razliku od prethodnog primjera, različite značke su u ovakvom označavanju u različitim kontekstima mogle biti označene različitim oznakama. U tablici 6.2 može se vidjeti da je maksimalni broj oznaka kojima je ista riječ označivana u ručnom korpusu veći. Jedna je riječ u pet različitih konteksta označena s pet različitih oznaka. Radi se o točki koja može poprimiti puno različitih funkcija:

- kraj rečenice,
- kraj rečenice u upravnom govoru,
- točka vezana za redni broj

i sl. Zbog ograničenosti označenog korpusa značke koje imaju veći broj mogućih oznaka pojavile su se u vrlo malenom broju. U tablicama 6.1 i 6.2 može se vidjeti da lematizacijska komponenta i ručno označeni korpus prate otprilike istu raspodjelu.

Broj mogućih oznaka	Broj znački	Udio u korpusu [%]
1	8180	98,7326
2	101	1,2191
3	1	0,0121
4	2	0,0241
5	1	0,0121

Tablica 6.2: Raspodjela riječi s više mogućih oznaka (prema ručno označenom korpusu)

Razina nesigurnosti	Broj znački	Udio u korpusu [%]
0	20343	93,7119
1	829	3,8188
2	256	1,1729
3	195	0,8983
4	68	0,3132
5	15	0,0691
6	2	0,0092

Tablica 6.3: Razina nesigurnosti u ručno dodijeljene oznake

Uvjerljivo najveći udio imaju riječi s jednom oznakom. Nešto manje je onih s dvije, a oznaka s tri i više različite oznake je zanemarivo malo. Uzrok razlike među raspodjelama vjerojatno se nalazi u činjenici da je u mjerenju za tablicu 6.1 korišteno označavanje koje istoj znački uvijek, neovisno o kontekstu, pridaje istu oznaku.

Bez formalne lingvističke poduke ljudskog označivača na nekim se značkama u korpusu veći broj oznaka činio potencijalno ispravnim. Te su značke označene jednom oznakom kojoj je dodana mjera nesigurnosti u tu oznaku. Ovakva vrlo subjektivna mjera nesigurnosti kretala se od 0 do 6. Udjeli oznaka s raznim razinama nesigurnosti mogu se vidjeti u tablici 6.3. Mjerenje je izvršeno na cijelom označenom korpusu.

6.2. Korištene mjere uspješnosti

U najvećem dijelu radova ne koristi se veliki broj mjera. Najčešće se koristi ili samo mjera postotka točnosti (engl. *accuracy*) definirana s

$$t = \frac{\text{broj točno označenih znački}}{\text{ukupni broj znački}} \cdot 100[\%], \tag{6.1}$$

ili mjera postotka pogreške dana s

$$p = 100 - t[\%]. \tag{6.2}$$

Je li oznaka neke značke *točna*, utvrđuje se usporedbom oznake koju je automatizirani označivač dodijelio nekoj riječi s oznakom koju joj je dodijelio ljudski označivač.

Uporaba preciznosti, odziva i mjere F_β uglavnom se preskače. Tek ponegdje, kao u (Agić et al., 2009), provodi se detaljnija analiza rezultata s osvrtom na vrijednosti navedenih mjera za svaku vrstu riječi, tj. svaku oznaku posebno.

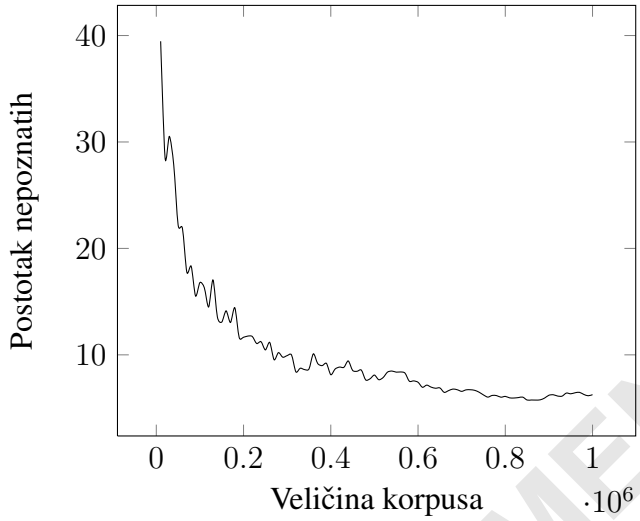
6.3. Pokusi i rezultati

U ovom su potpoglavlju prikazani najvažniji rezultati cjelokupnog rada – ocjena uspješnosti razvijenog automatiziranog označivača. Potpoglavlje također sadrži i raspravu o rezultatima i komentar.

6.3.1. Analiza udjela nepoznatih riječi u korpusu

Označivač razvijen u ovom radu, kao i svi označivači vrsta riječi razvijeni u okviru različitih znanstvenih radova, svoje najlošije rezultate ostvaruje na označavanju nepoznatih riječi. Velik udio nepoznatih riječi u dijelu korpusa na kojem se provodi ispitavanje uspješnosti nužno povlači i lošije rezultate označavanja. Stoga je uputno razmotriti kako se kreću udjeli nepoznatih riječi u ovisnosti o veličini korištenog korpusa. Analiza čiji se rezultati mogu vidjeti na slici 6.1 provedena je na većem dijelu Vjesnikovog korpusa pri čemu su korištene redom sve rečenice iz svih članaka i svih kategorija, počevši od najstarijeg dostupnog izdanja prema novijima. Slika 6.1 prikazuje utjecaj povećavanja veličine korpusa na udio nepoznatih riječi. Za svaku je veličinu korpusa 90% teksta uzeto za učenje, a 10% za testiranje. Nepoznate riječi su one riječi koje su se pojavile u 10% za testiranje, a nisu se pojavile u onih 90% za učenje.

Može se primijetiti da se za slučaj vrlo velikog korpusa (veličine 1 milijun znački) udio nepoznatih riječi smanjuje do oko 5.7%. U (Agić i Tadić, 2006) slično je mjerenje



Slika 6.1: Ovisnost veličine korištenog korpusa i udjela nepoznatih riječi

provedeno na puno manjem korpusu. Korišten je korpus CW100 koji sadrži 100000 znački iz članaka časopisa *Croatia Weekly*. Pri tome je za korpus veličine 100000, dakle od oko 10000 znački za testiranje njih tek 4.51% bilo nepoznato. Ovaj podatak indicira da je moguće da autori tekstova u korpusu CW100 u pisanju koriste nešto manji rječnik.

Kao što se moglo i očekivati, u tablici se vidi da se povećanjem veličine korpusa značajno smanjuje udio nepoznatih riječi u korpusu. Uzevši u obzir konačnost skupa riječi i njihovih oblika u hrvatskom jeziku, logičan je slijed da će se povećanjem korpusa povećati i šansa da se u njemu nađu sve moguće riječi ili bar njihova velika većina.

6.3.2. Ostvarena uspješnost automatiziranog označivača

U prosuđivanju ukupne uspješnosti označivača treba uzeti u obzir nepovoljan omjer poznatih i nepoznatih riječi u korpusu za testiranje. Označeni korpus je podijeljen na skup za učenje veličine 15000 znački i skup za testiranje veličine 5000 znački. Skup za testiranje sadržavao je samo 65,53% poznatih znački i čak 34,47% nepoznatih. Uzrok velikom udjelu nepoznatih riječi, kao što to govori graf na slici 6.1, je veličina označenog korpusa od samo 20000 znački. U radovima opisanim u poglavlju 2 korišteni su uglavnom veći korpusi s udjelom nepoznatih riječi manjim od 10%.

Tablica 6.4 prikazuje uspješnosti označivača mijenjajući veličinu skupa za učenje. Veličina korištenog dijela označenog korpusa varirana je, no u svim se mjerenjima u tablici koristio isti skup oznaka, OZNAKE2.

	Točnost [%]			
Dio korpusa [%]	Poznate	Nepoznate	Ukupno	Udio nepoznatih [%]
10	96,60	77,13	86,32	53,53
20	96,83	79,59	88,48	48,46
30	96,74	81,11	89,56	45,94
40	96,95	81,56	90,35	42,85
50	97,00	82,26	90,92	41,24
60	96,96	82,54	91,26	39,51
70	97,00	82,50	91,49	38,01
80	96,88	82,56	91,61	36,84
90	96,83	83,17	91,98	35,47
100	96,93	83,56	92,33	35,47

Tablica 6.4: Promjena uspješnosti označavanja u ovisnosti o veličini korpusa za učenje

	Točnost [%]		
Skup oznaka	Poznate	Nepoznate	Ukupno
OZNAKE1	95,08	68,88	86,05
OZNAKE2	96,93	83,56	92,33
OZNAKE-RED	98,00	83,40	92,97

Tablica 6.5: Utjecaj skupa oznaka na uspješnost označivača

Rezultati testova koji koriste prvotni skup oznaka – OZNAKE1, modificirani skup OZNAKE 2 i pojednostavljeni skup oznaka OZNAKE-RED prikazani su u tablici 6.5. Mjerenja su obavljena na cijelom dostupnom označenom korpusu, naravno, uz korištenje zaglađivanja vjerojatnosti u HMM-u drugog reda.

Usporedba označivača temeljenog na skrivenom Markovljevom modelu drugog stupnja s onim prvog stupnja vidi se u tablici 6.6. Mjerenja u tablici 6.6 provedena su uz korištenje cijelog označenog korpusa i skupa oznaka OZNAKE2.

Možda je iznenađujuća vrlo malena razlika između uspješnosti bigramskog (HMM prvog stupnja) i trigramskog (HMM drugog stupnja) označivača.

U tablici 6.7 prikazana je analiza točnosti rastavljena na pojedine vrste riječi, najzastupljenijih prema rjeđim vrstama. Može se vidjeti da su pogreške na nekim vrstama

	Točnost [%]		
Označivač	Poznate	Nepoznate	Ukupno
Bigramski	97,34	82,63	92,27
Trigramski	96,93	83,56	92,33

Tablica 6.6: Razlika između uspješnosti HMM-a prvog i drugog reda

riječi češće nego na drugima. Tako se puno grešaka pojavljuje na riječima iz stranih jezika (oznaka RSJ), što se moglo i predvidjeti, ako se uzme u obzir činjenica da će riječi iz stranih jezika vrlo često biti nepoznate riječi čija se oznaka predviđa na temelju sufiksa riječi u hrvatskom jeziku.

Od zastupljenijih riječi najviše je pogrešaka napravljeno na pridjevima i glagolima. Jedan uzrok slabijih rezultata na pridjevima vjerojatno se može pronaći u nezanemarivoj količini istih značaka koje su, ponekad u ovisnosti o širem kontekstu od onog koji razmatra HMM drugog reda, u nekim slučajevima pridjev (oznaka D) a u nekima glagolski pridjevi trpni (oznaka GDT). Slutnju da su ove dvije vrste riječi povezane podupire i slab rezultat na glagolskim pridjevima trpnim.

Točnost označavanja interpunkcije poput točaka, upitnika i uskličnika koji inače označavaju kraj rečenice, no u situacijama poput rečenica upravnog govora ne označavaju stvarni kraj rečenice (oznaka PKN), vrlo su niske. Takav rezultat potiče na drugačiji pristup detekciji takve interpunkcije.

Konačno, može se reći da označivač postiže vrlo dobre rezultate koji bi se mogli dodatno poboljšati uvođenjem posebnih obrazaca postupanja za neke vrste riječi, na primjer složene skraćenice (SS), riječi iz stranih riječi (RSJ), tzv. ne-riječi (NR), redne brojeve (BR) i jednostavne skraćenice (SJ), na kojima su označivač trenutno postiže niske točnosti.

Oznaka	Točnost [%]	Udio u skupu za testiranje [%]
I	95,48	1501/1572
D	82,18	498/606
J	98,11	570/581
V	94,40	337/357
Z	97,57	321/329
PZAREZ	100	290/290
GPOM	100	281/281
L	90	207/230
G	81,19	164/202
GDR	95,02	191/201
BG	94,56	139/147
GI	94,44	85/90
PNAVOD	97,50	78/80
GDT	65,22	30/46
SS	65,91	29/44
RSJ	18,42	7/38
Č	86,49	32/37
PTR	100	31/31
BR	70,83	17/24
PCRTA	100	15/15
PO	78,57	11/14
PKN	27,27	3/11
PGO	100	11/11
PGZ	100	11/11
GLS	100	9/9
NR	0	0/7
SJ	25	1/4
GLP	100	3/3
GIM	0	0/3
URI	0	0/3

Tablica 6.7: Točnosti na pojedinim vrstama riječi

7. Zaključak

Označavanje vrsta riječi je postupak u kojem se svim značkama u danom tekstu pridjeljuje oznaka koja sadrži informaciju o kategoriji riječi kojoj dana značka pripada. Poznavanje vrsta riječi u tekstu vrlo je korisno u brojnim postupcima na višim razinama obrade prirodnog jezika poput kategorizacije teksta, dohvata informacija, analize mišljenja i stava autora teksta, sintaktičke i semantičke analizu teksta ili pak strojnog prevođenja tekstova.

U okviru ovog diplomskog rada razvijen je i iskušan označivač vrsta riječi s razinama točnosti usporedivim s trenutnim najboljim ostvarenjima. Problem s kojim se često susreću istraživači koji rade na računalnoj obradi relativno manjih, u smislu broja govornika, prirodnih jezika – nepostojanje označenih korpusa nad kojim se mogu provoditi pokusi – nadvladan je tako što je stvoren ne prevelik, no značajan korpus s označenim vrstama riječi. Korpus će biti stavljen na raspolaganje drugim istraživačima kako bi eventualno mogao pomoći u budućim radovima. Označen je osnovnim oznakama vrsta riječi, što je dobra početna pozicija za eventualnu kasniju nadogradnju jednostavnim povećanjem njegove veličine i proširenjem skupa oznaka, po mogućnosti do potpunih morfosintaktičkih opisnika (engl. *morphosyntactic descriptor* – MSD).

Kao što je to slučaj u svim ostalim radovima u području označavanja vrsta riječi, nešto slabiji rezultati ostvareni su u označavanju nepoznatih riječi. Očekivana je i uobičajena pojava postizanja boljih rezultata na poznatim riječima nego nepoznatim. Međutim, može se spomenuti da su ovdje rezultati na nepoznatim riječima bili nešto slabiji od očekivanih. Označivač čije je učenje provedeno na 15000 znački a testiranje na 5000, postiže točnost od 83,64% na nepoznatim riječima, uz ukupnu konačnu točnost od 92,33%. Uz veći korpus za učenje, time i manji udio nepoznatih riječi, moguće je postizanje još boljih rezultata označivača.

Istraživanje označavanja vrsta riječi u tekstovima na hrvatskom jeziku moglo bi se nastaviti u nekom od više mogućih smjerova. Uz već predloženi rad na razvoju označenog korpusa, koji bi se u tom slučaju mogao proširiti oznakama više razine koje bi sadržavale sintaktičke informacije, grupirane imenske i glagolske fraze i, primjerice,

precizno označene granice surečenica i umetnutih rečenica, nastavak rada mogao bi se provesti i razvojem i iskušavanjem označivača zasnovanih na drugim pristupima. Označivači zasnovani na SVM-ima i maksimalnoj entropiji trenutno postižu najbolje rezultate dokumentirane za engleski jezik. U obzir dolazi i daljnja razrada pristupa zasnovanog na skrivenom Markovljevom modelu, svojevrstan hibridni pristup koji bi na model istreniran nadgledano dodao komponentu za nenadgledano učenje proširujući tako skup riječi poznatih označivaču unatoč korpusu ograničene veličine.

LITERATURA

M.M. Adler. *Hebrew morphological disambiguation: An unsupervised stochastic word-based approach*. Doktorska disertacija, Citeseer, 2007.

Ž. Agić i M. Tadić. Evaluating morphosyntactic tagging of croatian texts. U *Proceedings of the Fifth International Conference on Language Resources and Evaluation. ELRA, Genoa–Paris*, 2006.

Ž. Agić, Z. Dovedan, i M. Tadić. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 32(4):445–451, 2008. ISSN 0350-5596.

Ž. Agić, Ž. Dovedan, i M. Tadić. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica: An International Journal of Computing and Informatics*, 33(2):161–167, 2009. ISSN 0350-5596.

S.M. Aji i R.J. McEliece. The generalized distributive law. *Information Theory, IEEE Transactions on*, 46(2):325–343, 2002. ISSN 0018-9448.

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, i Marija Znika. *Hrvatska gramatika*. Školska knjiga, Zagreb, 2005.

T. Brants. TnT: a statistical part-of-speech tagger. U *Proceedings of the sixth conference on Applied natural language processing*, stranice 224–231. Association for Computational Linguistics, 2000.

E. Brill. A simple rule-based part of speech tagger. U *Proceedings of the third conference on Applied natural language processing*, stranice 152–155. Association for Computational Linguistics, 1992.

E. Brill. Some advances in transformation-based part of speech tagging. *Arxiv preprint cmp-lg/9406010*, 1994.

E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995. ISSN 0891-2017.

K.W. Church. A stochastic parts program and noun phrase parser for unrestricted text. U *Proceedings of the second conference on Applied natural language processing*, stranice 136–143. Association for Computational Linguistics, 1988.

C. Cortes i V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. ISSN 0885-6125.

D. Cutting, J. Kupiec, J. Pedersen, i P. Sibun. A practical part-of-speech tagger. U *Proceedings of the third conference on Applied natural language processing*, stranice 133–140. Association for Computational Linguistics, 1992.

W. Daelemans, J. Zavrel, P. Berck, i S. Gillis. MBT: A memory-based part of speech tagger generator. U *Proceedings of the Fourth Workshop on Very Large Corpora*, stranice 14–27, 1996.

S.J. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39, 1988. ISSN 0891-2017.

L. Dimitrova, N. Ide, V. Petkevic, T. Erjavec, H.J. Kaalep, i D. Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. U *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, stranice 315–319. Association for Computational Linguistics, 1998.

H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, i V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, stranice 155–161, 1997. ISSN 1049-5258.

T. Erjavec i B. Sárossy. Morphosyntactic tagging of Slovene legal language. *Informatica*, 30(4):483–488, 2006.

R.R. Forsythe. Morphological Analysis Toolbox. 2008.

J. Giménez i L. Màrquez. Fast and accurate part-of-speech tagging: The SVM approach revisited. *Recent advances in natural language processing III: selected papers from RANLP 2003*, stranica 153, 2003.

J. Giménez i L. Marquez. SVMTool: A general POS tagger generator based on Support Vector Machines. 2004.

Y. Goldberg, M. Adler, i M. Elhadad. EM can find pretty good HMM POS-taggers (when given a good start). U *Proc. of ACL*. Citeseer, 2008.

S. Goldwater i T. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. U *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, svezak 45, stranica 744, 2007.

I.J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963. ISSN 0003-4851.

B.B. Greene i G.M. Rubin. *Automatic grammatical tagging of English*. Dept. of Linguistics, Brown University, 1971.

J. Hajič i B. Hladká. Probabilistic and rule-based tagger of an inflective language: a comparison. U *Proceedings of the fifth conference on Applied natural language processing*, stranice 111–118. Association for Computational Linguistics, 1997.

Sanda Ham. *Školska gramatika hrvatskoga jezika*. Školska knjiga, Zagreb, 2002.

W.J. Hutchins. The Georgetown-IBM experiment demonstrated in January 1954. *Machine Translation: From Real Users to Research*, stranice 102–114, 2004.

N. Ide i J. Véronis. MULTEXT: Multilingual text tools and corpora. U *Proceedings of the 15th conference on Computational linguistics-Volume 1*, stranice 588–592. Association for Computational Linguistics, 1994.

E.T. Jaynes. Information theory and statistical mechanics. II. *Physical review*, 108(2): 171–190, 1957. ISSN 0031-899X.

S. Klein i R.F. Simmons. A computational approach to grammatical coding of English words. *Journal of the ACM (JACM)*, 10(3):334–347, 1963. ISSN 0004-5411.

K. Koskenniemi. Finite-state parsing and disambiguation. U *Proceedings of the 13th conference on Computational linguistics-Volume 2*, stranice 229–232. Association for Computational Linguistics, 1990.

H. Kucera i W.N. Francis. *Computational analysis of present-day American English*. Brown University Press Providence, RI, 1967.

S.E. Levinson, L.R. Rabiner, i M.M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, 1983.

A.V. Lukashin i M. Borodovsky. GeneMark. hmm: new solutions for gene finding. *Nucleic acids research*, 26(4):1107, 1998. ISSN 0305-1048.

C.D. Manning, H. Schütze, i MITCogNet. *Foundations of statistical natural language processing*, svezak 59. MIT Press, 1999.

Peter Matthews. *Morphology*. Cambridge University Press, second izdanju, 2009.

B. Merialdo. Tagging English text with a probabilistic model. *Computational linguistics*, 20(2):155–171, 1994. ISSN 0891-2017.

T. Nakagawa i Y. Matsumoto. Detecting errors in corpora using support vector machines. U *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, stranice 1–7. Association for Computational Linguistics, 2002.

T. Nakagawa, T. Kudoh, i Y. Matsumoto. Unknown word guessing and part-of-speech tagging using support vector machines. U *Proceedings of the sixth natural language processing pacific rim symposium*, stranice 325–331. Citeseer, 2001.

M. Poel, L. Stegeman, i R. Den Akker. A support vector machine approach to dutch part-of-speech tagging. U *Proceedings of the 7th international conference on Intelligent data analysis*, stranice 274–283. Springer-Verlag, 2007.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. ISSN 0018-9219.

Andrew Radford, Martin Atkinson, David Britain, Harald Clahsen, i Andrew Spencer. *Linguistics: An introduction*. Cambridge University Press, Zagreb, second izdanju, 2009.

Dragutin Raguž. *Praktična hrvatska gramatika*. Medicinska naklada, Zagreb, 1997.

A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. *IRCS Report*, stranice 97–08, 1997.

A. Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. U *Proceedings of the conference on empirical methods in natural language processing*, svezak 1, stranice 133–142, 1996.

C. Samuelsson. Morphological tagging based entirely on Bayesian inference. U *9th Nordic Conference on Computational Linguistics NODALIDA*, svezak 93, 1993.

Jan Šnajder. *Postupci morfološke normalizacije u pretrazivanju informacija i klasifikaciji teksta*. Doktorska disertacija, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2010.

S.M. Thede i M.P. Harper. A second-order hidden Markov model for part-of-speech tagging. U *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, stranice 175–182. Association for Computational Linguistics, 1999.

K. Toutanova i C.D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. U *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, stranice 63–70. Association for Computational Linguistics, 2000.

K. Toutanova, D. Klein, C.D. Manning, i Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. U *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, stranice 173–180. Association for Computational Linguistics, 2003.

Alan Turing. Computing Machinery and Intelligence, 1950.

J. Zavrel i W. Daelemans. Recent advances in memory-based part-of-speech tagging. U *VI Simposio Internacional de Comunicacion Social*, stranice 590–597. Citeseer, 1999.

Dodatak A
Skup oznaka OZNAKE1

Br.	Oznaka	Objašnjenje i primjer
1.	IO	opća imenica
		vlak, knjiga, pisanje, kraj, mercedes, kalodont
2.	IV	vlastita imenica
		Kolmogorov, Vatroslav
3.	G	glagol; obuhvaća prezent, aorist i imperfekt
		pišem, dođoste, dolaziste, bih, htjedoh
4.	GI	glagol u infinitivu
		hodati, penjati, nositi
5.	GIM	glagol u imperativu
		trčite, skoči, trče (bez čestice *neka*)
6.	GDR	glagolski pridjev radni
		hodao, rascvao, bila, htjeli, bio
7.	GDT	glagolski pridjev trpni
		nošen, ispitan, uznesen
8.	GLS	glagolski prilog sadašnji
		vozeći, hodajući, pišući
9.	GLP	glagolski prilog prošli
		došavši, diplomiravši
10.	D	opisni pridjev
		plava, pametna, željezni, brezova (metla), Šimunova (kuća)
11.	DL	pridjev koji nalikuje na prilog
		industrijski najrazvijenih zemalja
12.	Z	zamjenica

		što, ja, tebe, se
13.	ZU	upitna zamjenica tko, što, koji, čiji, kakav, kolik[1]
14.	BG	glavni (kardinalni) broj jedan, petsto, 24
15.	BR	redni (ordinalni) broj trideseti, drugi
16.	L	prilog vojnički (hodao), ponekad, nekamo, donekle, stoga
17.	J	prijedlog na, za, do, između
18.	V	veznik i, pa, da, kako
19.	SJ	jednostavna skraćenica itd., tj., prof.
20.	SS	složena skraćenica NATO, INA, GNU, SAD
21.	Č	čestica da, ne, neka
22.	U	usklik ih, oj, o
23.	PCP	interpunkcija – početak citata npr. znak „u „Ja sam svoju dionicu odradio “
24.	PCK	interpunkcija – kraj citata npr. znak “u „Sretno u Novu godinu! “
25.	PNO	interpunkcija – otvoreni navodnik prvi znak " u "divovski" ili znak „u „divovski “
26.	PNZ	interpunkcija – zatvoreni navodnik drugi znak " u "divovski" ili znak “u „divovski “
27.	PZO	interpunkcija – odvajajući zarez znak , u ...izvadi deset centa, a ostatak ostavi u novčaniku...

[1]Upitne su oblikom često identične npr. odnosnim zamjenicama, no pojavljuju se isključivo u upitnim rečenicama.

28.	PZN	interpunkcija – nabrajajući zarez znak , u ...tri jabuke, dvije šljive, lubenicu i krafnu...
29.	PGO	interpunkcija – otvorena zagrada
30.	PGZ	interpunkcija – zatvorena zagrada
31.	PKD	interpunkcija koja završava rečenicu
32.	PKN	znak koji inače završava rečenicu, ali ovdje ne Točke, uskličnici i upitnici u rečenicama upravnog govora
33.	PN	interpunkcija - zagrada u službi stavke nabrajanja Poredak je bio: 1.) Kostelić 2.) Palander
34.	PS	interpunkcija – spojnica manje-više
35.	PC	interpunkcija – crta (odv. rečenica, rasponi brojeva) Taj je posao – kažu – težak.
36.	PTR	interpunkcija – točka koja označava redni broj npr. točka u 1. Kostelić 2. Kirsch
37.	PI	interpunkcija – izostavnik
38.	PO	interpunkcija – ostalo ; / %
39.	NR	tzv. *ne-riječ* H_2O
40.	M	simbol simboli za valute, npr. simbol za euro: €
41.	RSJ	riječ iz stranog jezika (ne uključuje tuđice, usvojenice, prilagođenice itd.) npr. *buzzword* ali ne i *šou*
42.	URI	internetske poveznice *www.fer.hr* ali i pojedinačno: *www*, ., *google*, *com*

43.	?	oznaka nesigurnosti (djeluje kao sufiks) dodaje se na oznaku za koju ljudski označivač nije bio siguran

Dodatak B
Skup oznaka OZNAKE2

Br.	Oznaka	Objašnjenje i primjer
1.	I	imenica (i opće i vlastite)
		vlak, knjiga, pisanje, kraj, mercedes, kalodont
		Kolmogorov, Vatroslav
2.-3.	G (+POM)	glagol; obuhvaća prezent, aorist i imperfekt
		pišem, dođoste, dolaziste, bih, htjedoh
4.-5.	GI (+POM)	glagol u infinitivu
		hodati, penjati, nositi
6.-7.	GIM (+POM)	glagol u imperativu
		trčite, skoči, trče (bez čestice *neka*)
8.-9.	GDR (+POM)	glagolski pridjev radni
		hodao, rascvao, bila, htjeli, bio
10.-11.	GDT (+POM)	glagolski pridjev trpni
		nošen, ispitan, uznesen
12.-13.	GLS (+POM)	glagolski prilog sadašnji
		vozeći, hodajući, pišući
14.-15.	GLP (+POM)	glagolski prilog prošli
		došavši, diplomiravši
16.	D	opisni pridjev
		plava, pametna, željezni, brezova (metla), Šimunova (kuća)
17.	Z	zamjenica
		što, ja, tebe, se
18.	BG	glavni (kardinalni) broj
		jedan, petsto, 24

19.	BR	redni (ordinalni) broj trideseti, drugi
20.	L	prilog vojnički (hodao), ponekad, nekamo, donekle, stoga
21.	J	prijedlog na, za, do, između
22.	V	veznik i, pa, da, kako
23.	SJ	jednostavna skraćenica itd., tj., prof.
24.	SS	složena skraćenica NATO, INA, GNU, SAD
25.	Č	čestica da, ne, neka
26.	U	usklik ih, oj, o
27.	PNAVOD	znak navođenja znakovi „“" ’ « »
28.	PZAREZ	interpunkcija – zarez
29.	PGO	interpunkcija – otvorena zagrada
30.	PGZ	interpunkcija – zatvorena zagrada
31.	PKD	interpunkcija koja završava rečenicu
32.	PKN	znak koji inače završava rečenicu, ali ovdje ne Točke, uskličnici i upitnici u rečenicama upravnog govora
33.	PN	interpunkcija – zagrada u službi stavke nabrajanja Poredak je bio: 1.) Kostelić 2.) Palander
34.	PCRTA	interpunkcija – crta ili spojnica manje-više ili u rečenici *Taj je posao – kažu – težak.*

35.	PTR	interpunkcija – točka koja označava redni broj npr. točka u 1. Kostelić 2. Kirsch
36.	PI	interpunkcija – izostavnik
37.	PO	interpunkcija – ostalo ; / %
38.	NR	tzv. *ne-riječ* H_2O
39.	M	simbol simboli za valute, npr. simbol za euro: €
40.	RSJ	riječ iz stranog jezika (ne uključuje tuđice, usvojenice, prilagođenice itd.) npr. *buzzword* ali ne i *šou*
41.	URI	internetske poveznice *www.fer.hr* ali i pojedinačno: *www*, ., *google*, *com*
42.	?	oznaka nesigurnosti (djeluje kao sufiks) dodaje se na oznaku za koju ljudski označivač nije bio siguran

Dodatak C
Skup oznaka OZNAKE-RED

Fiksna oznaka	Riječi koje se njome uvijek označavaju
L (prilog)	upravo, zasad, ponovno, zacijelo upravo, zasad, ponovno, zacijelo, ipak, itekako, van, sasvim, svakako (Č), zajedno, naravno (rijetko D), trenutačno (D), širom, gdje, tamo, ondje, ponegdje, negdje, svugdje, doduše, potom, inače, ionako, dalje (D), nadalje, primjerice, zapravo, uskoro, odnosno (rijetko D), ubrzo, otkad, dnevno (rijetko D), također, pogotovo, onda (V), opet, tada, stoga, još, zato, dokle, ionako, tek, čak (Č), baš (Č), ujutro, uvijek, uvelike, unutra, lani, zatim, posve, nekoliko, godišnje (rijetko D), ponajprije (L), dosad, dosada (IO), širom (IO), štoviše, daleko (D), danas, tako L, uglavnom, napokon, puno (često D), malo (često D), odmah, ranije (rijetko D), nikad, nikada, otkud, donekle, sad (IO), sada (IO), međutim, toliko, oduvijek, dvaput (BG), uopće (Č), vrlo (rijetko D), kad (V kada je poslije zareza), već (V kada je poslije zareza), tu, gotovo (D), poglavito (D), uostalom (Č), barem (Č), samo (V ako je poslije zareza, D, Č)

	očito (D, Č), kako (V), osim (J, V), kao (V), prije (J ako je prije imenice u genitivu), poslije (J ako je prije imenice u genitivu), više (J ako je u smislu „iznad“, D)
J (prijedlog)	protiv (L), povodom (IO), potkraj (L), blizu (L), s, za, bez, od, preko, u, uz, na (rijetko U), uoči, tijekom (rijetko IO), iz, po, prema, iznad, o, nakon, zbog, među (IO), između, pod (često IO), kroz, poput, nad, izvan, unatoč, oko (često IO), pred, usred, pri, kod (često IO), nad, radi (često G), uslijed, umjesto (L s veznicima „da“i „što“), do (L)
V (veznik)	te, no, pošto (L), pa (Č), niti, čim (često Z), iako, ako, ali, pak, dok, premda, a, jer, ili, nego, ni (Č kao intenzifikator)
Č (čestica)	ama, eto
IO (opća imenica)	posto (L)
BG (glavni broj)	pol (IO)

Označavanje vrste riječi u tekstovima na hrvatskome jeziku

Sažetak

Označavanje vrsta riječi važna je predradnja u brojnim područjima istraživanja obrade prirodnog jezika. Počevši od obrade rečenica na sintaktičkoj razini, preko trenutno vrlo aktualnih radova u analizi stavova i mišljenja autorâ tekstôva, do strojnog prevođenja, informacija o vrsti svake od riječi u tekstu vrlo je korisna. U okviru ovog diplomskog rada dan je pregled dosadašnjih radova u označavanju vrsta riječi s osvrtom na rezultate različitih pristupa programskom ostvarenju automatiziranog označivača za brojne svjetske jezike, kao i uvod u vezanu lingvističku problematiku. Programski je ostvaren automatizirani označivač zasnovan na skrivenim Markovljevim modelima te su komentirani postignuti rezultati na hrvatskom jeziku – 92,33% na ograničenom dostupnom korpusu označenom u okviru rada.

Ključne riječi: označavanje vrste riječi, hrvatski jezik, skriven Markovljev model, nadzirano strojno ucenje, obrada prirodnog jezika, računalna lingvistika

Tagging parts of speech in Croatian texts

Abstract

Part of speech tagging is an important early step in many research areas in natural language processing. Beginning with syntactic analysis of sentences, through currently very fashionable areas of opinion and sentiment analysis, up to machine translation, having part of speech information is very useful. Within this Master's thesis an overview of related work in different languages and using different approaches to tagging is given along with a detailed description of the underlying linguistic intricacies of the Croatian language. An HMM-based tagger was implemented and its results – 92.33% on the limited available corpus hand-tagged for the purpose of this thesis – were documented and analysed.

Keywords: part of speech tagging, croatian language, hidden Markov model, supervised machine learning, natural language processing, computational linguistics