TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4272

MODEL KOHERENTNOSTI TEKSTOVA NA HRVATSKOM JEZIKU TEMELJEN NA ANALIZI ENTITETA

Jura Šlosel

Zagreb, lipanj 2015.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2015.

ZAVRŠNI ZADATAK br. 4272

Pristupnik: **Jura Šlosel (0036464311)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Model koherentnosti tekstova na hrvatskome jeziku temeljen na analizi entiteta**

Opis zadatka:

Koherentnost je bitna diskursna značajka teksta budući da izravno utječe na njegovu razumljivost. Računalno modeliranje i analiza koherentnosti teksta u domeni je diskursne analize, sve značajnijeg područja u okviru obrade prirodnog jezika. Tipične su primjene automatsko generiranje teksta, sažimanje dokumenata te ocjenjivanje eseja. U literaturi postoji niz teorija tzv. lokalne koherentnosti, od kojih se mnoge temelje na načinu spominjanja diskursnih entiteta u tekstu, te su predloženi odgovarajući računalni modeli koji vrlo uspješno određuju stupanj koherentnosti teksta.
U okviru završnoga rada potrebno je proučiti pristupe za računalno modeliranje i računalnu analizu koherentnosti s naglaskom na pristupe za modeliranje lokalne koherentnosti. Proučiti pristupe temeljene na analizi entiteta, posebice model temeljen na rešetci entiteta (engl. entity grid) Barzilay i Lapate (2008). Razviti programsku implementaciju tog postupka i primijeniti ga na tekstove na hrvatskome jeziku, odnosno na postojeće novinske korpuse s označenim diskursnim entitetima. Razmotriti proširenja modela predložena u literaturi. Primijeniti model na zadatak određivanja poretka rečenice te razmotriti mogućnost primjene modela na zadatak određivanja razumljivosti teksta. Provesti iscrpno vrednovanje, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 12. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Računalna obrada prirodnog jezika (engl. *Natural Language Processing, NLP*) je brzo razvijajuća grana računarske znanosti i umjetne inteligencije. NLP ima široku primjenu danas kad količina znanja, koje je uglavnom reprezentirano tekstualnim dokumentima, raste eksponencijalnom brzinom i izvan je sposobnosti ljudi da se s time nose bez pomoći računala. Neki od zadataka NLP-a su razumijevanje prirodnog jezika, raspoznavanje relacija između entiteta u tekstu i vstrojno generiranje teksta. Danas postoje mnogi programski sustavi za strojno generiranje teksta, primjerice sustavi za strojno prevođenje, sustavi za generiranje teksta iz koncepta te sustavi za automatsko sažimanje teksta. Ti su sustavi bitna karika za približavanje te goleme količine podataka ljudima odnosno za interakciju čovjeka i računala.

Cilj kojem se teži pri strojnom generiranju teksta je da rezultat bude nemoguće razaznati od teksta kakvog bi napisao čovjek. Jedno od bitnih obilježja takvog teksta jest koherentnost, stupanj povezanosti činjenica, koncepata, ideja, rečenica – sastavnih jedinica teksta – u koherentnu cjelinu. Pri strojnom generiranju teksta bitno je moći automatski ocjeniti obilježja generiranog teksta, uključujući koherentnost. Ocjenjivanje koherentnosti pročitanog zadatak je vrlo lagan i prirodan ljudima, koji su kroz iskustvo naučili prepoznati nekoherentan tekst. Točnije, ljudi će podsvjesno, bez razmišljanja nekoherentan tekst razumjeti teško ili ga uopće neće razumjeti. Zbog toga se kao algoritamsko rješenje problema određivanja koherentnosti prirodno javlja strojno učenje karakteristika koherentnih odnosno nekoherentnih tekstova.

U ovom radu predložen je model koherentnosti tekstova na hrvatskom jeziku temeljen na analizi entiteta. Predloženi model reprezentira tekst rešetkom diskursnih entiteta odakle se vade vjerojatnosti pojave različitih prijelaza uloge entiteta u susjednim rečenicama. Nadziranim učenjem na skupu manje i više koherentih tekstova model uči koji su prijelazi česti za koherentne odnosno nekoherentne tekstove. Model je posve automatiziran i ne zahtijeva ručno označavanje dokumenata, već se automatski stvara skup sintetički označenih dokumenata. Rezultantni model kao ulaz prima niz tekstova i rangira ih po stupnju koherentnosti. Model je vrednovan na standardnom

zadatku vrednovanja koherentnosti, zadatku određivanja poretka rečenice. Model opisan u ovom radu temelji se na radu Barzilay i Lapata (2008), koji su prvi opisali ovaj model.

Ostatak rada strukturiran je na sljedeći način. Drugo poglavlje daje pregled teorija koherentnosti i računalnih modela koherentnosti. U trećem poglavlju detaljno je opisan model rešetke entiteta te alati koji se koriste pri izradi rešetke entiteta - stroj s potpornim vektorima (engl. *Support Vector Machine, SVM*) i ovisnosni parser za hrvatski jezik. Četvrto poglavlje daje kratak pregled strukture programskog rješenja koje je implementacija modela. U petom poglavlju opisani su svi eksperimenti provedeni s modelom te analiza rezultata. Rad je zaključen u šestom poglavlju.

2. Problem: Analiza koherentnosti

2.1. Definicija koherentnosti

Tekst je koherentan ako je organiziran na način da njegovi djelovi slijede prirodno jedan iz drugog, odnosno svaki dio je jasno smješten u kontekst ostalih djelova, a nijedan dio nije smješten tako da daje informaciju čiji smisao nije jasan u pripadnom kontekstu. Pojam "dio teksta" ovdje nije dobro definiran. Može se raditi o rečenici, nekoliko uzastopnih rečenica, u duljim tekstovima može se raditi o paragrafu ili čak poglavlju u članku ili knjizi. Da bi tekst bio koherentan, on bi trebao zadovoljavati navedeno obilježje na svim razinama.

Neke teorije koherentnosti koncentrirane su samo na koherentnost na niskoj razini, na prijelaze među susjednim rečenicama. To su teorije lokalne koherentnosti. Teorije globalne koherentnosti proučavaju pravila koherentnosti na razini čitavog teksta.

Navedimo dvije definicije koherentnosti iz literature. Mann i Thompson (1988) definira koherentnost ovako: [1]

Definicija 1 *Jedna formulacija koherentnosti jest odsutnost* non sequitur [2] *i rupa. Odnosno, za svaki dio koherentnog teksta postoji neka funkcija, neki valjan i čitatelju jasan razlog za njegovu prisutnost. Nadalje, ne osjeća se da u tekstu nedostaju neki djelovi.*

Document Understanding Conference (DUC), konferencija istraživača sažimanja tekstova, definira sljedeća obilježja sažetaka po kojima će se ocjenjivati njihova kvaliteta: gramatičnost, neredundantnost, jasnoća referenciranja, fokus i koherentnost. Činjenica da je koherentnost odvojena od tih srodnih obilježja teksta daje bolji uvid u to što koherentnost obuhvaća, a što ne. Također treba primjetiti da se definicija 1 odnosi na koherentnost sažetaka, što su kratki tekstovi (DUC traži da duljina bude u prosjeku

[1] Preuzeto s `http://www.sfu.ca/rst/01intro/intro.html`

[2] Latinski za "ne slijedi iz"

250 znakova) te se ne odnosi nužno na koherentnost dugih tekstova. DUC daje sljedeći opis za svojstvo sažetaka "struktura i koherentnost". [3]

Definicija 2 *Sažetak treba biti dobro strukturiran i dobro organiziran. Sažetak ne smije biti samo hrpa nepovezanih informacija, nego treba iz rečenice u rečenicu rasti u koherentno tijelo informacija o temi.*

Zanimljivo je primijetiti da za kratak tekst to može značiti da mu je koherentnost definirana jedino poretkom rečenica. Odnosno, za skup rečenica kratkog teksta različiti poretci rečenica bit će različite koherentnosti. Valja zapamtiti ovu primjedbu jer će voditi način razmišljanja kod vrednovanja modela rešetke entiteta.

Za razumijevanje koherentnosti korisno je vidjeti primjere nekoherentih tekstova. Nekoliko primjera koherentnih i nekoherentnih tekstova slijedi u nastavku.

Primjer 1 *Ustav ne daje predsjedniku izravno takvu moć. No ipak, predsjednik ima dužnost ne prekršiti Ustav. Pitanje je je li njegovo jedino oružje veto.*[4]

Druga rečenica daje kontrast prvoj i objašnjenje za iduću rečenicu. Bilo koji drugi poredak tih triju rečenica bio bi nekoherentan.

Primjer 2 *Svi se slažu da većinu starih mostova u državi treba ili popraviti ili zamijeniti. No ima neslaganja o načinu kako to učiniti.* [5]

I ovdje druga rečenica daje kontrast prvoj. Ako se rečenice zamijene bez preoblikovanja sadržaja, rezultat je nekoherentan.

Primjeri 3 i 4 prikazuju nekoherentnost ne dvije ili tri rečenice, već kratkog teksta.

Primjer 3 *[Otprilike 30% tinejdžera postane eksperimentalnim pušačima.] [Znamo da svaki dan 3000 tinejdžera počne pušiti.] [Otprilike 90% njih je jednom vjerovalo da nikad neće pušiti.] [Od onih koji počnu pušiti, otprilike 90% će nastaviti nositi kutiju cigareta i upaljač do kraja života.] [Bez obzira koliko tko htio ostati nepušačem,] [činjenica je da je pritisak vršanjaka na pušenje u srednjoj školi veći nego u bilo kojem drugom životnom razdoblju.] [Otprilike 75% adolescenata u nekom trenutku će podići cigaretu i dopustiti znatiželji da preuzme.]* [6]

[3] Preuzeto s `http://duc.nist.gov/duc2007/quality-questions.txt`

[4] Preuzeto iz Lin et al. (2011)

[5] Preuzeto iz Lin et al. (2011)

[6] Preuzeto iz Marcu (1996)

Za tekst iz primjera 3 možda nije odmah posve jasno zašto je nekoherentan. Problem se može opisati rekavši da "tekst nije zadovoljavajuć jer je potrebno potruditi se da bi ga se razumjelo". [7] Koherentnija verzija teksta dana je u primjeru 4.

Primjer 4 *[Bez obzira koliko tko htio ostati nepušačem,] [činjenica je da je pritisak vršanjaka na pušenje u srednjoj školi veći nego u bilo kojem drugom životnom razdoblju:] [Otprilike 75% adolescenata u nekom trenutku će podići cigaretu i dopustiti znatiželji da preuzme.] [Otprilike 30% tinejdžera postane eksperimentalnim pušačima.] [Od onih koji počnu pušiti, otprilike 90% će nastaviti nositi kutiju cigareta i upaljač do kraja života.] [Znamo da svaki dan 3000 tinejdžera počne pušiti,] [iako je činjenica da je otprilike 90% njih jednom vjerovalo da nikad neće pušiti.]* [8]

2.2. Motivacija: strojno generiranje teksta

Primjeri protočnih sustava za strojno generiranja teksta su automatsko sažimanje, odgovaranje na pitanja i generiranje teksta iz koncepta. Protočni sustavi za strojno generiranje teksta koriste jedan od sljedećih dvaju općenitih pristupa. Tradicionalnim pristupom generira se jedan visoko kvalitetan izlaz, za stvaranje kojeg u sustavu postoji mnogo ograničenja (engl. *constraints*), koja su nerijetko protuslovna. Drugi je pristup u dvije faze - *generiraj-poredaj*. U prvoj fazi generira se nekoliko mogućih izlaza, različitih zbog varijacija parametara s obzirom na različita ograničenja, a u drugoj fazi ti se izlazi poredaju funkcijom rangiranja.

Pristup generiraj-poredaj omogućava veću fleksibilnost i može se koristiti u svim fazama protočnog sustava za strojno generiranje teksta. Naime, protočni sustavi za strojno generiranje teksta su višeslojni, a tek posljednja komponenta, sustav za planiranje teksta, kao izlaz daje konačan tekst, dok ostale komponente generiraju međustrukture npr. diskursno stablo za planiranje teksta.

U sustavima generiraj-poredaj potrebna je funkcija za poredavanje rezultata s obzirom na određeno obilježje. Komponente za planiranje teksta tad često koriste funkciju za rangiranje po koherentnosti. Upravo to je glavna motivacija za izradu modela koji je tema ovog rada i razlog zašto taj model funkcionira na način da kao ulaz prima nekoliko tekstova, a kao izlaz daje njihov poredak po koherentnosti.

[7] Marcu (1996)

[8] Preuzeto iz Marcu (1996)

2.3. Teorije koherentnosti

Teorije koherentnosti generalno se dijele u dvije skupine:

1. teorije lokalne koherentnosti
2. teorije globalne koherentnosti

Lokalna koherentnost teksta odnosi se na koherentnost na razini nekoliko susjednih rečenica - na načinu na koji tema prelazi iz jedne rečenice u sljedeću, pa rečenicu iza i tako dalje. Globalna koherentnost teksta odnosi se na koherentnost čitavog teksta kao cjeline. Tekst može imati visoku lokalnu koherentnost no globalno biti nekoherentan. Razlog tome je što se lokalna koherentnost ne odnosi na tekst u cjelini već samo na prijelaze i povezanost susjednih rečenica, a ostalo ne uzima u obzir. Tekst visoke lokalne koherentnosti može npr. započinjati ili završavati na nespretan način, dati loš uvod odnosno zaključak za cjelinu ili biti loše obrađen u sredini i tako ukupno biti loše povezan.

Odlomci 2.3.1, 2.3.2 i 2.3.3 opisuju neke od teorija globalne odnosno lokalne koherentnosti. Teorije lokalne koherentnosti temeljene na analizi entiteta su temelj modela opisanog u ovom radu i obrađene su u poglavlju 2.4.

2.3.1. Teorija retoričke strukture

Teorija retoričke strukture (Mann i Thompson, 1988) (engl. *Rhetorical Structure Theory, RST*) razvijena je sredinom 80-ih u sklopu istraživanja strojnog generiranja teksta s ciljem stvaranja teorije strukture diskursa potrebne za modele generiranja teksta. RST proučava opis strukture tekstova gradivnim blokovima na dvije razine. Prva razina su relacije jezgri (engl. *nucleus*) i satelita, a druga razina su sheme. RST se primjenjuje na modeliranje koherentnosti proučavanjem koji su uzorci tih gradivnih blokova uočljivi u koherentnim tekstovima. U nastavku je ukratko objašnjen sistem relacija jezgri i satelita.

Najčešći strukturni obrazac jest da su dva dijela teksta povezana tako da jedan dio ima određenu ulogu s obzirom na drugi. Kaže se da između njih postoji određena relacija. Ta dva dijela teksta nazivaju se jezgra ("važniji" od dvaju djelova) i satelit. Ovdje "dio teksta" nije definiran i može se raditi o svakoj razini teksta: podrečenice (zavisne ili nezavisne rečenice unutar složene rečenice), rečenica i niz rečenica.

Primjeri relacija:[9]

- dokaz
    - jezgra: tvrdnja
    - satelit: informacije namjenjene povećanju čitateljevom vjerovanju u tvrdnju
    - primjer: "J(Koliko god bilo primamljivo, ne smijemo se latiti svakog popularnog problema na koji naiđemo.) S(Kad to činimo, trošimo vrijedne, ograničene resurse, dok za to vrijeme drugi igrači sa superiornim resursima rade prikladniji posao.)"
- koncesija
    - jezgra: situacija koju autor potvrđuje
    - satelit: druga, naizgled nekonzistentna situacija, no autor povrđuje i nju
    - primjer: "S(Koliko god bilo primamljivo,) J(ne smijemo se latiti svakog popularnog problema na koji naiđemo.)"
- kontrast (primjer relacije koja nema jezgru i satelit već dvije jezgre)
    - prva jezgra: jedna alternativa
    - druga jezgra: druga alternativa
    - primjer: "J(Zbog nedostatka laktaze, većina odraslih ljude ne može probaviti mlijeko.) J(U populacijama koje piju mlijeko, odrasli ljudi imaju više laktaze, što je vjerojatno posljedica prirodnog odabira.)"

RST ima mnogo relacija, ovo su samo neki od primjera.

Dva dijela teksta mogu imati svoju relaciju jezgra-satelit i ponovo biti jezgra ili satelit u relaciji s trećim dijelom teksta. Takva rekurzivna analiza teksta – njegova raščlamba na djelove među kojima vrijede relacije – je razlog zašto se teorija retoričkih struktura u računalnim modelima koherentnosti najčešće koristi za modeliranje globalne koherentnosti.

Analiza teksta nije jednoznačna – moguće su različite analize istog teksta. Povezanost dijelova teksta može se drugačije shvatiti, pa tako i drugačije opisati.

2.3.2. Modeli temeljeni na sadržaju

Za razliku od teorije retoričke strukture, koja karakterizira diskurs retoričkim relacijama, elementima neovisnima o domeni, postoje modeli koncentrirani na jednako

[9] Primjeri su preuzeti s `http://www.sfu.ca/rst/01intro/definitions.html`. Oznake "J" i "S" u primjerima znače "jezgra" i "satelit".

fundamentalnu dimenziju strukture diskursa koja jest ovisna o domeni - sadržaj. Pojam *sadržaj* diskursa odgovara temi i promjenama teme u tekstu. Modeli za opisivanje sadržaja diskursa trebaju znati prepoznati da npr. tekstovi o potresima tipično sadrže informacije o jačini potresa, lokaciji i broju žrtava te da izvještaj o broju žrtava tipično dolazi prije informacija o pokušajima spašavanja. Takvi modeli su ovisni o domeni jer su uzorci sadržaja općenito, u svim domenama odjednom, prevarijabilni da bi ih bilo moguće prepoznati i opisivati računalnim modelima.

Model Barzilay i Lee (2004) je primjer modela diskursa temeljnog na sadržaju. Njihov model bazira se na skrivenim Markovljevim modelima koji opisuju prijelaze u temi kroz tekst iz vjerojatnosne perspektive. Stanja njihovog skrivenog Markovljevog modela odgovaraju vrstama informacije karakterističnima toj domeni, a prijelazi u stanjima odgovaraju mogućim poretcima iznošenja tih informacija u danoj domeni.

2.3.3. Latentna semantička analiza

Latentna semantička analiza (engl. *Latent Semantic Analysis, LSA*) je tehnika analiziranja odnosa među dokumentima i među pojmovima koje ti dokumenti sadrže stvaranjem skupa koncepata povezanih s dokumentima i pojmovima. Dokument se predstavlja vektorom čiji su elementi pojmovi u tom dokumentu. LSA pretpostavlja da se riječi sličnog značenja pojavljuju u sličnim djelovima teksta. LSA se koristi za računanje semantičke povezanosti dvaju dokumenata, tako što se semantička povezanost računa kao kosinus kuta njihovih vektora.

Foltz i Landauer (1998) primjenjuju LSA na računanje lokalne koherentnosti. Umjesto čitavog dokumenta, svaki vektor predstavlja jednu rečenicu. Ukupna lokalna koherentnost teksta zatim se računa kao prosječna semantička povezanost između svake dvije susjedne rečenice u tekstu.

Ovaj način računanja koherentnosti posve je automatiziran, ne zahtijeva ručno označavanje podataka, kao i model rešetke entiteta. Za razliku od modela rešetke entiteta, ovaj je model leksikaliziran, odnosno sadrži i ovisi o uzorcima riječi u tekstovima.

2.4. Pristupi lokalnoj koherentnosti temeljeni na analizi entiteta

U teorijama koherentnosti, entitet je bilo kakva stvar, osoba, pojava, apstrakcija ili bilo što drugo što može biti tema teksta. Entitet se često definira kao skup koreferen-

tih imenskih skupova (engl. *noun phrase*) u tekstu, gdje "koreferentni" znači da se ti imenski skupovi odnose (engl. *to refer*) na isti entitet. Drugim riječima, jedan entitet može se spomenuti više puta u tekstu i načini spominjanja mogu biti različiti, no sva spominjanja odnose se na isti entitet. Primjerice, ako se radi o osobi, entitet se može prvi put realizirati imenom, prezimenom i atributima, drugi put samo imenom, a treći put zamjenicom.

Teorije lokalne koherentnosti temeljene na analizi entiteta imaju dugu tradiciju u lingvistici koja seže do 70-ih. Zajednička pretpostavka različitih teorija jest da je koherentnost teksta usko povezana s načinom na koji se entiteti uvode u tekst i u njemu raspravljaju. Ovo se razmatranje često formalizira ograničavanjem lingvističke realizacije i raspodjele entiteta u koherentnom tekstu. Pod "lingvistička realizacija" misli se na izraze kojima je entitet spomenut u tekstu: puno ime osobe, puna imenica, zamjenica.

Svim teorijama je zajednička pretpostavka da su kroz tekst različiti entiteti različito istaknuti (engl. *salient*) i da za istaknute entitete vrijede drugačija pravila pojavljivanja nego za neistaknute. U teoriji centriranja (engl. *Centering Theory*) (Grosz i Joshi, 1995), istaknutost entiteta određena je upravo realizacijom entiteta. Što je viši stupanj anaforizacije[10] entiteta, to se entitet smatra istaknutijim. Razmišljanje iza toga jest da entiteti koji su slabo specificirani u tekstu, odnosno spominje ih se zamjenicama, moraju biti važni jer je u tekstu implicitno pretpostavljeno da se zna o kojem entitetu se radi. Na istaknutost entiteta također utječe gramatička uloga entiteta – važniji entiteti imat će važniju gramatičku ulogu poput subjekta i objekta. Neke teorije istaknutost smatraju binarnom relacijom, a neke skalarnom.

Teorije definiraju koherentnost teksta pomoću raspodjele entiteta kroz djelove diskursa, razlikujući između istaknutih i neistaknutih entiteta. Intuitivno, koherentni tekstovi sastojat će se od mnogo susjednih djelova s istim entitetima, bez naglih skokova s jedne teme na drugu, odnosno s jedne skupine entiteta na sasvim drugu. Time se postiže kontinuiranost teme. Teorija centriranja formalizira promjene u kontinuiranosti teme pomoću prijelaza između susjednih djelova teksta. Prijelazi imaju različit rang, s obzirom na stupanj koherentnosti.

[10] Anafora je uporaba izraza čija interpretacija ovisi o drugom izrazu u tekstu, koji je prethodio ili koji slijedi dani izraz. U rečenici "Ako vidiš *Anu*, pozdravi je.", izraz *Anu* je prethodnik izraza *je*. Oba izraza odnose se na isti entitet. Bez spomena imena *Ana*, izraz *je* ne može se točno interpretirati. U tekstovima duljine nekoliko rečenica, prethodnik ili sljedbenik izraza te izraz mogu biti u različitim rečenicama. U tekstu se tada javljaju realizacije entiteta koje su vrlo općenite te njihova interpretacija zahtjeva prethodnika ili sljedbenika. Ako entitet ima mnogo realizacija koje su takvi izrazi, on ima visok stupanj anaforizacije.

3. Model rešetke entiteta

3.1. Rešetka entiteta

Rešetka entiteta je dvodimenzionalno polje s *m* redaka i *n* stupaca, gdje je *m* broj rečenica, a *n* broj diskursnih entiteta u tekstu. U ovom radu svaka pojedina imenica je jedan diskursni entitet, a sva pojavljivanja iste imenice odnose se na isti entitet. Dakle, rešetka entiteta teksta sadrži onoliko diskursnih entiteta koliko je u tekstu različitih imenica. U svakoj ćeliji rešetke s koordinatama *(i,j)* piše gramatička uloga koju *j*-ti entitet ima u rečenici *i*. Moguće gramatičke uloge su subjekt (oznaka "S"), objekt (oznaka "O") te niti subjekt niti objekt (oznaka "X"). Ako se *j*-ti entitet ne javlja u *i*-toj rečenici, u rešetci *(i,j)* to je označeno oznakom "-". U rešetci entiteta svakom entitetu može biti pridružena maksimalno jedna uloga po rečenici, pa se u slučaju da se entitet pojavljuje više puta u rečenici za ulogu uzima uloga najveća po važnosti od svih gramatičkih uloga koju entitet ima u toj rečenici. Važnost je pritom definirana na sljedeći način: $S > O > X$. Dakle, jednom tekstu odgovara rešetka entiteta čiji retci odgovaraju rečenicama u tekstu, a stupci odgovaraju entitetima poredanima po redoslijedu njihovog prvog pojavljivanja u tekstu.

Valja primjetiti da je u modelu rešetke entiteta kao središnja jedinica analize odabrana rečenica. U literaturi ne postoji konsenzus koja bi se jezična jedinica trebala koristiti kao središnja. Jedna druga mogućnost jest raspodijeliti složene rečenice na podrečenice (zavisne i nezavisne podrečenice) i uzeti njih za središnju jedinicu. Time bi bio izbjenut kompromis da se po rečenici jednom entitetu može dati samo jedna gramatička uloga. No, za razliku od alata za izdvajanje rečenica, nema dovoljno preciznih alata za tu svrhu. Također, takav model bio bi složeniji.

U tablici 3.1 dan je primjer rešetke entiteta izrađene za tekst iz tablice 3.2.

Tablice 3.1 i 3.2 služe prikazivanju načina izgradnje rešetki entiteta iz teksta te stoga ne sadrže uobičajene pogreške iz predobrade teksta. U odlomku 3.4 opisan je utjecaj pogreški predobrade na rad modela.

Valja primijetiti da je entitetu *učenik* u prvom retku rešetke pridružena uloga X,

Tablica 3.1: Rešetka entiteta za tekst iz tablice 3.2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	X	X	X	S	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	O	O	X	O	X	X	S	X	X	-	-	-	-	-	-	-	-
3	-	-	-	-	-	X	-	-	-	-	-	-	-	-	S	O	X	S	-	-	-	-
4	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	X	X	X

Tablica 3.2: Kratak novinski članak s označenim imenicama tj. diskursnim entitetima. Brojevi uz entitete označavaju redni broj entiteta u tekstu, odnosno redni broj stupca u rešetci entiteta koji odgovara tom entitetu.

1	U [četvrtak]$_1$ poslijepodne u prometnoj [nesreći]$_2$ teško je stradao [učenik]$_3$ [M.S.]$_4$ (10) pri [izlasku]$_5$ iz [autobusa]$_6$.
2	Naime, on je, izašavši iz [autobusa]$_6$, pretrčavao [cestu]$_7$ na pješačkom [prijelazu]$_8$ iza zaustavljenog [autobusa]$_6$, kad je na njega [automobilom]$_9$ [marke]$_{10}$ » [Lada]$_{11}$ [Niva]$_{12}$ « naletio [R.M.]$_{13}$ (33) iz [Mirkovca]$_{14}$.
3	[Vozač]$_{15}$ nije, kao što je trebao, zaustavio [vozilo]$_{16}$ za [vrijeme]$_{17}$ dok su [djeca]$_{18}$ izlazila iz [autobusa]$_6$.
4	Liječnička [pomoć]$_{19}$ teško ozlijeđenom [M.S.]$_4$ pružena je u [Klinici]$_{20}$ za dječje [bolesti]$_{21}$ u [Zagrebu]$_{22}$.

iako je *učenik* podčvor subjekta *M.S*. Druga je mogućnost pridružiti svim imenicama ulogu koju ima njihov korijenski čvor. U tom bi slučaju *učenik* dobio ulogu S. Implikacija ovog odabira na model jest to što se svim apozicijama pridružuje uloga X te model raspolaže s manje uzoraka s ulogama S i O.

3.2. Vektori značajki

Iz svake rešetke entiteta gradi se jedan vektor značajki koji reprezentira tekst. Pri odabiru značajki općenito je potrebno učiniti kompromis između izražajnosti i računalne složenosti. Uz veću izražajnost, vektori su složeniji i veće su dimenzije te pritom raste količina podataka potrebna za učenje modela. No veća izražajnost znači da je model "pametniji" i može raspoznati pravilnosti koje manje izražajan model ne bi mogao. Kako bi se obuzdala dimenzionalnost vektora značajki u modelu rešetke entiteta, koristi se vjerojatnosni model. Dodatna prednost vjerojatnosnog modela je što vrlo jednostavno jednako manipulira tekstovima različitih duljina.

Značajke su vjerojatnosti pojave različitih vrsta *prijelaza* gramatičke uloge entiteta

u susjednim rečenicama. Prijelaz je definiran kao niz $\{S,O,X,-\}^n$ i on predstavlja promjenu gramatičke uloge jednog entiteta u *n* susjednih rečenica. Npr. za entitet "autobus" iz tablice 3.1 prijelaz duljine 4 je [X,O,X,-].

Vjerojatnost pojave prijelaza duljine *n* definira se kao ukupan broj pojavljivanja tog prijelaza u rešetci entiteta podjeljen s brojem svih prijelaza duljine *n*. Za prethodni primjer, pripadna značajka prijelaza [X,O,X,-] je 0.02727 (broj pojavljivanja tog prijelaza u rešetci entiteta je 1, broj svih prijelaza duljine 4 je 22).

Prema saznanjima teorija koherentnosti temeljenih na analizi entiteta, koherentni tekstovi trebali bi biti takvi da susjedne rečenice sadrže slične entitete. Očekuje se da većina entiteta bude uvedena u tekst, obrađena u malom broju susjednih rečenica i kasnije se više ne spominje. Očekuje se da postoji manji broj *istaknutih* entiteta koji imaju više spominjanja kroz čitav tekst od neistaknutih te često imaju bitnu gramatičku ulogu. Ukoliko se to prevede u model rešekte entiteta, od koherentnih tekstova očekuje se da imaju malen broj gustih stupaca i velik broj gotovo praznih stupaca, zatim da uloge entiteta u gustim stupcima često budu subjekt i objekt. Nekoherentni tekstovi u pravilu ne bi trebali imati ta svojstva.

3.3. Lingvističke dimenzije rešetke entiteta

Odabir izvora lingvističkog znanja koji su bitni za točnost modela i način njihove reprezentacije jedan su od središnjih problema u stvaranju modela koherentnosti temeljenih na analizi entiteta. Prethodni pristupi uglavnom se slažu u tome koja obilježja distribucije entiteta su relevantna, no postoji neslaganje o načinima kako tim obilježjima izgraditi model.

Pri odabiru parametara za model rešetke entiteta, traži se kompromis između sljedećih stavki:

1. Lingvistička važnost parametra

2. Točnost parametra pri njegovom automatskom izračunu

   Naime, zamisao ovog modela je da može raditi posve automatski, bez potrebe za ljudskim označavanjem podataka za treniranje i bez da ljudi obrađuju ulazne tekstove označavajući informacije relevantne za rešetku entiteta.

3. Veličina rezultantnog prostora značajki.

3.3.1. Vađenje diskursnih entiteta

Već je rečeno da je u ovom radu svaka pojedina imenica jedan diskursni entitet, a sva pojavljivanja iste imenice odnose se na isti entitet. Grupiranje imenica po identitetu u entitete programski je implementirano uspoređivanjem lema dviju imenica. Ukoliko su im leme jednake, dvije imenice spadaju pod isti entitet. Za vađenje vrste riječi (engl. *part-of-speech tag*) i leme riječi u tekstu koriste se HunPos tagger i CST lemmatiser s modelima za hrvatski jezik opisanima u (Agić et al. 2013).

U originalnom radu (Barzilay i Lapata, 2008) definiran je kao "skup koreferentnih imenskih skupova (engl. *noun phrase*)" u tekstu. Ta definicija odgovara definiciji diskursnih entiteta u teorijama lokalne koherentnosti temeljene na analizi entiteta. Ona omogućava da se isti entitet u tekstu realizira i punim imenom i djelom imena i osobnom zamjenicom itd. To je fleksibilnije, bliže prirodnom tekstu i točnije od pretpostavke da će realizacija entiteta u tekstu svaki put biti istom imenicom. No, još ne postoji javno dostupan alat za automatsku ekstrakciju koreferentnih imenskih skupova za hrvatski jezik. Štoviše, takav alat postoji za malen broj jezika. Jedan od njih je engleski i taj je alat primjenjen u modelu Barzilay i Lapata (2008). Točnost alata primjenjenog u osnovnom modelu je vrhunska (engl. *state of the art*), a ta vrhunska točnost iznosi 70.4 F-mjere na MUC-6 i 63.4 F-mjere na MUC-7. Niske performanse takvog alata su razlog zašto su Barzilay i Lapata (2008) eksperimentirali i s pojednostavljenim modelom koji se koristi u ovom radu, a koji ne koristi alat za ekstrakciju koreferentnih imenskih skupova. Isto su činili i radovi koji nadograđuju osnovni model.

Primjeri 5 i 6 pokazuju razliku između koreferencije temeljene na identitetu imenice i potpune koreferencije imenskih skupova. Na oba teksta entiteti su izvađeni ručno, ne automatskim alatima. Primjeri služe samo da pokažu spomenute razlike.

Primjer 5 *[Stribor]$_{e1}$ je [student]$_{e2}$ [Fakulteta elektrotehnike i računarstva]$_{e3}$. [On]$_{e1}$ [svaki dan]$_{e4}$ prolazi pored [knjižnice]$_{e5}$ na [putu]$_{e6}$ do [faksa]$_{e3}$.*

Imenica "Stribor" i zamjenica "on" svrstani su pod isti entitet *e1*, što i jesu. Isto tako i fraza "Fakultet elektrotehnike i računarstva" i imenica "faksa" pod entitet *e3*.

Primjer 6 *[Stribor]$_{e1}$ je [student]$_{e2}$ [Fakulteta]$_{e3}$ elektrotehnike]$_{e4}$ i [računarstva]$_{e5}$. On svaki [dan]$_{e6}$ prolazi pored [knjižnice]$_{e7}$ na [putu]$_{e8}$ do [faksa]$_{e9}$.*

Primjer 6 pokazuje kako se pri razrješavanju koreferencije identitetom imenica zamjenice ("on") uopće ne smatraju entitetima, a imenski skupovi koji prestavljaju jedan entitet ("Fakultet elektrotehnike i računarstva") rastavljaju se u onoliko entiteta koliko

imenica sadrže. Da je u u drugoj rečenici umjesto imenice "faks" iskorištena imenica "fakultet", ovaj model bi ih svrstao u isti entitet.

3.3.2. Gramatička funkcija

Nekoliko teorija koherentnosti temeljenih na analizi entiteta slaže se da gramatička uloga entita u tekstu ukazuje na važnost entiteta u tekstu. Većina teorija razlikuje između tri uloge: subjekt, objekt i preostale uloge. Pritom je važnost entiteta s obzirom na gramatičku ulogu rangirana na način da je subjet važniji od objekta, a objekt od preostalih uloga.

Za automatsko vađenje gramatičkih uloga riječi koristi se ovisnosni parser za hrvatski jezik, Agić i Merkler (2013).

Rešetka entiteta ograničena je na način da jedan entitet u jednoj rečenici može imati samo jednu gramatičku ulogu. U tekstu se naravno isti entitet može pojaviti više puta u različitim ulogama. No ovo je stvar kompromisa i jednostavnosti modela. Ako bi se omogućile dvije ili tri različite uloge entiteta po rečenici, dimenzionalnost vektora značajki također bi se povećala toliko puta, a i dalje ne bi bilo podržano rješenje za općenitu situaciju gdje se entitet u rečenici može pojaviti n puta s n različitih uloga.

3.3.3. Istaknutost entiteta

Teorija centriranja i druge teorije diskursa predlažu da način na koji je entitet uveden u tekst i kasnije realiziran u tekstu ovisi o globalnoj ulozi tog entiteta u tekstu. Po tome entiteti mogu imati različite stupnjeve istaknutosti u tekstu. U ovom modelu entitet može ili biti istaknut ili ne, odnosno podjela je binarna.

Neki modeli predlažu podjelu na više stupnjeva, no i ovdje je odluka donesena zbog kompromisa s veličinom prostora značajki. Naime, istaknutost entiteta ovdje je predstavljena tako što model odvojeno uči uzorke prijelaza za istaknute entitete od drugih. Odnosno, čitav vektor značajki podjeljen je na dva djela jednake duljine: prva polovica predstavlja vjerojatnosti određenih prijelaza za neistaknute entitete, a druga polovica vjerojatnosti istih prijalaza za istaknute entitete. Podjela entiteta na istaknute i neistaknute se uzima u obzir pri izračunu vjerojatnosti prijelaza. Ovo je dalje objašnjeno u poglavlju 3.3.4.

Istaknutost entiteta određena je frekvencijom pojavljivanja entiteta u tekstu. Ukoliko entitet ima barem dva pojavljivanja, smatra se istaknutim, u suprotnom neistaknutim.

3.3.4. Izgradnja vektora značajki

U tablicama 3.3 i 3.4 prikazan je dio vektora značajki za rešetku entiteta iz tablice 3.1 i za tekst iz tablice 3.2. Potpuni vektor značajki ima 160 značajki te zbog veličine nije prikazan u cjelosti. Umjesto toga prikazane su značajke 1.–16. (tablica 3.3) i 81.–96. (tablica 3.4). Značajke 1.–16. odnose se na vjerojatnosti prijelaza duljine 2 za neistaknute entitete, a značajke 81.–96. na vjerojatnosti prijelaza duljine 2 za istaknute entitete. Značajke 17.–80. reprezentiraju vjerojatnosti prijelaza duljine za neistaknute entitete, a značajke 97.–160. vjerojatnosti prijelaza duljine 3 za istaknute entitete. Postupak izračunavanja vjerojatnosti za prijelaze duljine 3 jednak je postupku za prijelaze duljine 2 te je dovoljno prikazati taj podskup da postupak bude jasan.

Razlog zašto je potrebno 16 značajki za sve vjerojatnosti prijelaza duljine 2 je to što je broj različitih prijelaza duljine n jednak broju permutacija niza duljine n čiji su članovi elemenati skupa {S,O,X,-} koji ima četiri elemenata. Takvih permutacija ima 4^n.

Tablica 3.3: Značajke 1.–16. vektora značajki za rešetku entiteta iz tablice 3.1. Vjerojatnosti su zaokružene na 3 decimale.

SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	–
0	0	0	0.067	0	0	0	0.033	0	0	0	0.167	0.083	0.033	0.150	0.467

Tablica 3.4: Značajke 81.–96. vektora značajki za rešetku entiteta iz tablice 3.1. Vjerojatnosti su zaokružene na 3 decimale.

SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	–
0	0	0	0.167	0	0	0.167	0	0	0.167	0	0.167	0	0	0.167	0.167

Značajka 4 u tablici 3.3 predstavlja vjerojatnost pojave prijelaza [S,-] za neistaknute entitete. Ona je jednaka kvocijentu broja pojavljivanja prijelaza [S,-] i ukupnog broja prijelaza duljine 2 za neistaknute entitete. Kvocijent iznosi 4/60 = 0.067. Treba primjetiti da je broj prijelaza duljine 2 jednak 60 jer se u obzir ne uzimaju entiteti 4 i 6 iz tablice 3.1, koji su istaknuti.

Značajka 4 u tablici 3.4 predstavlja vjerojatnost pojave prijelaza [S,-] za *istaknute* entitete. Ona je jednaka kvocijentu broja pojavljivanja prijelaza [S,-] i ukupnog broja prijelaza duljine 2 za *istaknute* entitete. Kvocijent iznosi 1/6 = 0.167. Broj prijelaza duljine 2 sad je 6 jer se u obzir uzimaju samo istaknuti entiteti 4 i 6.

3.4. Predobrada teksta

Svaki se tekst predobrađuje kako bi se izvadile gramatičke uloge diskursnih entiteta za igradnju rešetke entiteta. Za to se koristi ovisnosni parser za hrvatski jezik koji su razvijali Agić i Merkler (2013).

Prije parsanja ovisnosnim parserom, tekst se obrađuje alatom za segmentaciju teksta po rečenicama i tokenizaciju (podjelu rečenica u znakove, tj. riječi). Taj alat pružio je Laboratorij za analizu teksta i injženjerstvo znanja, TakeLab [1]. Dodatno, prije ulaza u ovisnosni parser, tekst se obrađuje alatima za određivanje vrste riječi (engl. *part-of-speech*) te alatom za lematizaciju. Ti su alati opisani u (Agić et al. 2013).

Ovisnosni parser otkriva sintaksnu strukturu svake rečenice teksta, odnosno način na koji su djelovi rečenice međusobno povezani te kako je rečenica izgrađena. Ovisnosni parser koristi ovisnosnu gramatiku za opis strukture rečenice. Takve su gramatike prikladne za opis jezika sa slobodnim poretkom riječi, kao što je hrvatski jezik. U ovisnosnoj gramatici svaka riječ ovisi o drugoj riječi koja je njen korijen, izuzev korijena rečenice, što je u pravilu glagol odnosno predikat rečenice. Izlaz ovisnosnog parsera sadrži, između ostalog, gramatičku ulogu svake riječi u rečenici, a to je ono što se koristi u modelu.

Jedan od osnovnih zahtjeva modela jest da bude potpuno automatiziran, bez potrebe za ljudskim označavanjem. Prednost takvog modela je veća brzina, skalabilnost i fleksibilnost u svakoj fazi rada sustava. Negativna strana jest šum koji parser unosi u sustav. U predobradi tekstova za ovaj model ulančana su 4 programska alata. Točnost tog parsera je 75–78%. U nastavku je dan izlaza parsera za primjer rečenice s ciljem ocrtavanja uobičajenih grešaka koje se događaju i koje treba uzeti u obzir pri analizi rezultata modela.

U primjeru 7 dana je rečenica i rezultat parsanja te rečenice. Za trenutna razmatranja bitna su polja: 1, 2, 4 i 8. Prvo polje je redni broj znaka u rečenici, gdje je znak ili riječ ili dijakritični znak. Drugo polje je znak. Četvrto polje je vrsta riječi. Osmo polje je gramatička uloga riječi s obzirom na korijensku riječ.

Primjer 7 *Naime, prema informacijama PU požeško-slavonske, automobilom marke »opel vectra«, registracijske oznake B-HK 9733 (D), upravljao je D.Č. (30) iz Berlina.*

[1] `http://takelab.fer.hr/`

1	*Naime*	*naime*	*N*	*N*	*-*	*25*	*Sb*	*-*	*-*
2	*,*	*,*	*Z*	*Z*	*-*	*3*	*Punc*	*-*	*-*
3	*prema*	*prema*	*S*	*S*	*-*	*1*	*Prep*	*-*	*-*
4	*informacijama*	*informacija*	*N*	*N*	*-*	*3*	*Obj*	*-*	*-*
5	*PU*	*pu*	*N*	*N*	*-*	*6*	*Ap*	*-*	*-*
6	*požeško*	*požeško*	*N*	*N*	*-*	*7*	*Sb*	*-*	*-*
7	*-*	*-*	*Z*	*Z*	*-*	*9*	*Punc*	*-*	*-*
8	*slavonske*	*slavonska*	*N*	*N*	*-*	*7*	*Atr*	*-*	*-*
9	*,*	*,*	*Z*	*Z*	*-*	*3*	*Punc*	*-*	*-*
10	*automobilom*	*automobil*	*N*	*N*	*-*	*9*	*Elp*	*-*	*-*
11	*marke*	*marka*	*N*	*N*	*-*	*10*	*Atr*	*-*	*-*
12	*»*	*»*	*Z*	*Z*	*-*	*10*	*Punc*	*-*	*-*
13	*opel*	*opel*	*N*	*N*	*-*	*6*	*Ap*	*-*	*-*
14	*vectra*	*vectar*	*N*	*N*	*-*	*13*	*Atr*	*-*	*-*
15	*«*	*«*	*Z*	*Z*	*-*	*13*	*Punc*	*-*	*-*
16	*,*	*,*	*Z*	*Z*	*-*	*13*	*Punc*	*-*	*-*
17	*registracijske*	*registracijski*	*A*	*A*	*-*	*18*	*Atr*	*-*	*-*
18	*oznake*	*oznaka*	*N*	*N*	*-*	*19*	*Atr*	*-*	*-*
19	*B-HK*	*b-hk*	*N*	*N*	*-*	*25*	*Sb*	*-*	*-*
20	*<num>*	*9733*	*M*	*M*	*-*	*19*	*Atr*	*-*	*-*
21	*(*	*(*	*Z*	*Z*	*-*	*22*	*Punc*	*-*	*-*
22	*D*	*d*	*M*	*M*	*-*	*25*	*Pred*	*-*	*-*
23	*)*	*)*	*Z*	*Z*	*-*	*22*	*Punc*	*-*	*-*
24	*,*	*,*	*Z*	*Z*	*-*	*25*	*Punc*	*-*	
25	*upravljao*	*upravljati*	*V*	*V*	*-*	*0*	*Pred*	*-*	*-*
26	*je*	*biti*	*V*	*V*	*-*	*25*	*Aux*	*-*	*-*
27	*D.Č*	*d.č*	*N*	*N*	*-*	*25*	*Sb*	*-*	*-*
28	*.*	*.*	*Z*	*Z*	*-*	*0*	*Punc*	*-*	*-*
29	*(*	*(*	*Z*	*Z*	*-*	*30*	*Punc*	*-*	*-*
30	*<num>*	*30*	*M*	*M*	*-*	*33*	*Atr*	*-*	*-*
31	*)*	*)*	*Z*	*Z*	*-*	*30*	*Punc*	*-*	*-*
32	*iz*	*iz*	*S*	*S*	*-*	*25*	*Prep*	*-*	*-*
33	*Berlina*	*berlin*	*N*	*N*	*-*	*32*	*Atr*	*-*	*-*
34	*.*	*.*	*Z*	*Z*	*-*	*0*	*Punc*	*-*	*-*

U modelu rešetke entiteta od interesa su samo imenice, dakle znakovi čije je 4. polje jednako "N" (engl. *noun*). U modelu je za imenice također bitno jesu li subjekt,

objekt ili niti prvo ni drugo. U ovoj rečenici, imenice kako ih je prepoznao parser su: *naime* (1. znak), *informacijama* (4. znak), *PU* (5. znak), *požeško* (6. znak), *slavonske* (8. znak), *automobilom* (10. znak), *marke* (11. znak), *opel* (13. znak), *vectra* (14. znak), *oznake* (18. znak), *B-HK* (19. znak) *D.Č.* (27. znak), *Berlina* (33. znak). Od navedenih znakova, njih tri uopće nisu imenice: *naime*, *požeško*, *slavonske*. Od imenica koje su zbilja imenice, parser daje znaku *B-HK* ulogu subjekta, no jedini pravi subjekt je *D.Č.*.

3.5. Stroj potpornih vektora

Strojevi potpornih vektora (engl. *Support Vector Machines, SVM*) su modeli nadziranog učenja za analizu podataka i raspoznavanje uzoraka koji se koriste za klasifikaciju uzoraka. Temeljem skupa uzoraka za učenje modela (skup vektora), SVM gradi skup hiperravnina u prostoru vektora takvih da svaka bude što udaljenija od susjednih točaka (uzoraka iz skupa za učenje) u prostoru. Udaljenosti između hiperravnina i točaka nazivaju se marginama. SVM pokušava maksimizirati marginu za svaku hiperravninu, slijedeći intuiciju da time pospješuje generalizaciju modela.

Istrenirani SVM za klasifikaciju uzoraka je funkcija čiji je ulaz vektor, a izlaz razred (cijeli broj) kojem pripada taj vektor. U ovom se radu koristi SVM za rangiranje, što je modifikacija modela za klasifikaciju. SVM za rangiranje umjesto razreda vektora ima za izlaz bodovanje (realan broj) vektora s ulaza. Poredavanjem vektora s obzirom na bodovanje dobiva se željeni rezultat: međusobni poredak, odnosno rang vektora. Yu et al. (2009) daje slijedeću formalnu definiciju problema SVMa za rangiranje.

Neka je R skup za treniranje SVM-a za rangiranje, $R = \{(\mathbf{x}_1,y_1),...,(\mathbf{x}_m,y_m)\}$. $\mathbf{x}_i$ je vektor u prostoru uzoraka, a y_i je rang (engl. *ranking*) vektora $\mathbf{x}_i$. y_i je cijeli broj, takav da vrijedi $\mathbf{x}_i > \mathbf{x}_j$ (što znači da je vektor $\mathbf{x}_i$ rangiran na višoj poziciji od vektora $\mathbf{x}_j$) kad vrijedi $y_i < y_j$. Zbog jednostavnosti, u ovoj definiciji se pretpostavlja da je poredak takav da nijedna dva vektora nemaju isti rang.

Cilj SVM-a za rangiranje je naučiti funkciju F koja zadovoljava $F(\mathbf{x}_i) > F(\mathbf{x}_j)$ za sve parove $\{(\mathbf{x}_i,\mathbf{x}_j) : y_i < y_j\}$ iz skupa za treniranje R te generalizira i izvan njega. Funkcija $F(\mathbf{x}_i)$ je funkcija koja vektoru $\mathbf{x}_i$ pridružuje njegov rang. Ako je F linearna funkcija, učenje funkcije F ekvivalentno je izračuvanju vektora težina $\mathbf{w}$ t.d. za većinu parova $\{(\mathbf{x}_i,\mathbf{x}_j) : y_i < y_j\}$ iz skupa R vrijedi:

$$F(\mathbf{x}_i) > F(\mathbf{x}_j) \implies \mathbf{w} \cdot \mathbf{x}_i > \mathbf{w} \cdot \mathbf{x}_j \implies F(\mathbf{x}_i > \mathbf{x}_j) > 0$$

Funkcija F nije uvijek linearna. To ovisi o raspodjeli točaka iz skupa za učenje u prostoru – razredi u koje su raspodjeljeni nisu uvijek linearno separabilni. Taj se pro-

blem zaobilazi preslikavanjem izvornog prostora značajki u novi visokodimenzionalni prostor, u kojem se linearna separacija može provesti jednostavnije i točnije. Pritom se skalarni umnošci vektora u visokodimenzionalnom prostoru definiraju pomoću jezgrenih funkcija (engl. *kernel function*), funkcija čiji su argumenti dva vektora iz izvornog prostora. Time se zaobilazi skupo (iz perspektive računalnih resursa) eksplicitno preslikavanje vektora u visokodimenzionalni prostor te taj prostor ostaje definiran samo implicitno.

4. Implementacija

Programsko rješenje je implementirano u Javi 1.7. Nisu korištene vanjske biblioteke već jedino JavaSE [1].

Rješenje se sastoji od nekoliko međusobno nezavisnih komponenti. Komponente međusobno surađuju spremajući rješenje u datoteku odakle ga iduća komponenta u protočnom sustavu učitava i koristi. Dva su razloga za takvo oblikovanje:

1. Program SVMrank (Joachims, 2006), koji se koristi u ovom radu, kao ulaz prima datoteku čiji su sadržaj vektori značajki.

2. Time što je izlaz iz svake komponente sustava tekstna datoteka ostavlja se mogućnost kasnijeg detaljnog proučavanja svakog koraka, što je korisno u istraživačkom radu.

Te nezavisne komponente obavljaju sljedeće funkcije:

- predobrada dokumenata, izvedeno u paketu `hr.fer.takelab.zr.slosel.nlptools`
- stvaranje permutacija dokumenata, izvedeno u paketu `hr.fer.takelab.zr.slosel.permutations`
- stvaranje rešetki entiteta, izvedeno u paketu `hr.fer.takelab.zr.slosel.entitygrid`
- stvaranje vektora značajki, izvedeno u paketu `hr.fer.takelab.zr.slosel.featurevector`
- treniranje i testiranje SVM modela, optimizacija SVM parametara, računanje točnosti modela, izvedeno u paketu `hr.fer.takelab.zr.slosel.svm`

[1] JavaSE (engl. *Java Platform, Standard Edition*) je programska platforma za razvoj aplikacija u programskom jeziku Java. Ona sadrži širok spektar osnovnih biblioteka za aplikacije za osobna računala i poslužitelje. Link: `http://www.oracle.com/technetwork/java/javase/overview/index.html`

– analiza rezultata eksperimenata, izvedeno u paketu `hr.fer.takelab.zr.slosel.statistics`

Komponente su međusobno nezavisne jer se mogu pozivati neovisno jedna o drugoj. U svakom od gore navedenih paketa postoji barem jedan razred s metodom `main`, što znači da se može pokrenuti kao zaseban program. Primjerice, za stvaranje vektora značajki iz postojećih datoteka s rešetkama entiteta, poziva se metoda `main` razreda `hr.fer.takelab.zr.slosel.featurevector.FeatureVectorsCreator`.

Postoje tri razreda koji logički povezuju preostale, inače nezavisne komponente:

– `hr.fer.takelab.zr.slosel.TrainPipeline` pruža metodu `main` koja automatizira ciklus treniranja modela: pretprocesiranje, permutiranje tekstova, stvaranje rešetki entiteta, stvaranje vektora značajki i treniranje SVM modela.

– `hr.fer.takelab.zr.slosel.RankPipeline` pruža metodu `main` koja automatizira ciklus testiranja modela: pretprocesiranje, permutiranje tekstova, stvaranje rešetki entiteta, stvaranje vektora značajki i testiranje SVM modela.

– `hr.fer.takelab.zr.slosel.ConfigProperties` jest jedinstveni objekt (engl. *singleton*) koji na jednom mjestu obuhvaća i upravlja svim promjenjivim parametrima sustava – primjerice ime datoteke gdje će se spremiti SVM model – i odakle im može pristupiti svaki razred.

Svi promjenjivi parametri sustava mogu se specificirati ili u konfiguracijskoj datoteci `config.properties`, ili prijenosom parametara u naredbenom retku pri pokretanju `main` metode bilo kojeg razreda.

5. Vrednovanje

5.1. Određivanje poretka rečenica teksta

Model rešetke entiteta vrednovan je na zadatku određivanja poretka rečenica teksta. To je jedan od standardnih načina testiranja koherentnosti. Zadatak se izvodi na sljedeći način. Svakom tekstu *T* pridruži se skup *P* – skup *n* različitih permutacija rečenica teksta *P*. Model zatim za svaki uređeni par *(T,P)* određuje koherentnost originalnog teksta i svih permutacija – *n+1* ocjena koherentnosti po uređenom paru. Ukupan broj permutacija za sve originalne tekstove neka je *N*. Ukupan broj slučajeva kad je pojedinoj permutaciji dodjeljena veća koherentnost nego odgovarajućem originalnom tekstu neka je *M*. Točnost modela definira se ovako:

$$tocnost = \frac{M}{N} \cdot 100\% \tag{5.1}$$

Iza ovakve definicije točnosti modela stoji pretpostavka da je svaki originalni tekst koherentniji od njemu pridruženih permutacija njegovih rečenica. Ta je pretpostavka povezana s definicijama koherentnosti tekstova iz poglavlja 2.1, posebice definicije 2 o koherentnosti sažetaka. U ovom radu permutacije tekstova stvaraju se automatski algoritmom slučajne zamjene dviju rečenica u tekstu za svaku rečenicu. Na taj je način vrlo lako stvoriti sintetički označen korpus – nije potrebno ručno označavati. Pritom se gubi na točnosti, no dobiva na brzini. Jednostavno i brzo se može stvoriti novi skup dokumenata za treniranje drugog modela, npr. kako bi se stvorio model za drugu tematsku domenu tekstova. Gubitak točnosti odnosi se na činjenicu da se ne preispituje pretpostavka da je bilo koja permutacija različita od originalnog teksta manje koherentna od originalnog teksta. Ta pretpostavka vjerojatno nije posve točna, odnosno postoje permutacije rečenica koje su koherentnije od originalnog teksta, ili je razliku u koherentnosti teško odrediti. Kad bi se dokumenti za ovaj zadatak označavali ručno, za treniranje i testiranje modela koristile bi se samo one permutacije koje ljudi procjenjuju manje koherentnima i izbjegla bi se pogreška navedene pretpostavke.

Pretpostavka o većoj koherentnosti originalnog teksta od bilo koje permutacije do-

voljno je dobra da opravda objašnjeni pristup. Razlog tome je što je način na koji ljudi pišu tekstove takav da ideje prevedu u niz promjena teme i zatim u niz rečenica. Postoji mnogo načina da se izvorne ideje prevedu u niz rečenica, no nakon što je bilo koji niz stvoren, on je koherentan (to je realno očekivanje za bilo koji tekst koji piše čovjek). Međutim, promjenom poretka rečenica koje su u danom poretku tvorile koherentnu cjelinu gube se poveznice između rečenica i prijelazi u temi koji su imali smisla za tekst samo u izvornom obliku.

5.2. Korpusi tekstova

Za treniranje modela odabran je korpus novinskih članaka crne kronike Vjesnika. To je podskup korpusa hrWaC (Ljubešić i Erjavec 2011), korpusa tekstova na hrvatskom jeziku skupljenih na Internetu s domene *.hr*.

Korpus crne kronike sastoji se od 23717 tekstova. Ovaj rad slijedi Barzilay i Lapata (2008) koji su pojedini SVM model trenirali na 100 tekstova. Odabrana su dva podskupa veličine 100 tekstova svaki, "Vjesnik-promet" i "Vjesnik-ubojstva" za treniranje modela. Obrazloženje za podjelu jest očekivanje da će tematski različite domene tekstova obilježavati različiti uzorci u prijelazima entiteta te da će se modeli do neke mjere razlikovati u naučenim pravilima. Također, u ove su skupove izdvojeni u prosjeku kraći tekstovi, za kakve je model lokalne koherentnosti prikladan. Korpusi Vjesnik-promet i Vjesnik-ubojstva imaju različite distribucije duljine tekstova, no slične prosječne duljine. Duljina teksta računa se kao broj rečenica u tekstu. Vjesnik-promet ima prosječnu duljinu 9.52, a Vjesnik-ubojstva 11.80 rečenica.

Potencijalni tekstovi za korpuse Vjesnik-promet i Vjesnik-ubojstva dobiveni su pretragom pojavljivanja ključnih riječi i izraza u čitavom korpusu. Za korpus Vjesnik-promet, tekstovi su morali sadržavati izraz "prometna nesreća" ili koji drugi oblik tog izraza u bilo kojem broju i padežu. Time je osigurano da se korpus sastoji od tekstova vrlo sličnog sadržaja. Za korpus Vjesnik-ubojstva, tekstovi su morali sadržavati bilo koji od nizova znakova "ubojstv", "ubojic", "umorstv". Dobiveni tekstovi pokazali su se raznolikijima od tekstova korpusa Vjesnik-promet te su izdvojeni tekstovi o pokušajima ubojstva i samoubojstvima i razmatrali su se samo tekstovi kojima to nije tema. Preciznije, odstranjeni su tekstovi koji su sadržavali nizove "pokušaj", "samoubojstv" ili "samoubojic".

Tako dobiveni korpus o prometu sadržavao je 2668 tekstova, a korpus o ubojstvima 2283. Iz korpusa o prometu izdvojeno je 200 tekstova, 100 za skup za treniranje (korpus Vjesnik-promet) i 100 za skup za testiranje (korpus test-Vjesnik-promet). Isto tako

za korpuse Vjesnik-ubojstva i test-Vjesnik-ubojstva. Ti su tekstovi odabrani tako da se duljinama poklapaju tekstovima u skupu za treniranje Barzilay i Lapata (2008).

Za čim bolju usporedbu s modelom Barzilay i Lapata (2008), napravljen je pokušaj da Vjesnik-promet i Vjesnik-ubojstva sadrže tekstove istih duljina kao korpusi u njihovom radu. Vjesnik-promet pritom odgovara korpusu *Earthquakes*, a Vjesnik-ubojstva odgovara korpusu *Airplanes*. Međutim, izračun duljina tekstova na engleskom jeziku napravljen je jednostavnom podjelom teksta u rečenice po točkama zbog čega je rezultat neprecizan te se korpusi Vjesnik-promet i Vjesnik-ubojstva ipak razlikuju u duljinama od korpusa originalnog modela. Korpusi za testiranje modela, test-Vjesnik-promet i test-Vjesnik-ubojstva, sadrže tekstove jednakih duljina kao odgovarajući korpusi za treniranje. Pripadna distribucija veličine tekstova dana je u tablicama 5.1 i 5.2.

Tablica 5.1: Korpus Vjesnik-promet i test-Vjesnik-promet. Broj tekstova s pripadnim brojem rečenica.

broj rečenica:	4	5	6	7	8	9	10	11	12	13	14	15	17	18	19	20	22	23	26
broj tekstova:	4	14	14	15	9	6	4	6	2	2	5	5	3	1	1	3	1	2	1

Tablica 5.2: Korpus Vjesnik-ubojstva i test-Vjesnik-ubojstva. Broj tekstova s pripadnim brojem rečenica.

broj rečenica:	6	7	8	9	10	11	12	13	14	15	16	17	18	19
broj tekstova:	1	1	7	16	15	15	8	8	6	13	3	3	2	2

U tablici 5.1 vidi se da je gotovo 50% tekstova u korpusima Vjesnik-promet i test-Vjesnik-promet duljine pet, šest ili sedam rečenica. Tekstova duljine veće ili jednake dvadeset rečenica ima 7%, dok u korpusima Vjesnik-ubojstva i test-Vjesnik-ubojstva (tablica 5.2) uopćena nema tekstova te duljine (najveća duljina je 19). Distribucija duljina tekstova tog korpusa bliža je zvonolikoj krivulji. Ovdje je 50% tekstova duljine devet, deset ili jedanaest rečenica. Od duljine 8 do 19, broj tekstova se monotono spušta, osim za tekstove duljine 15, kojih ima 13%.

5.3. Eksperimenti

U sklopu rada provedeno je nekoliko eksperimenata s modelom rešetke entiteta, s ciljem uvida u obilježja modela. Svi su eksperimenti varijacije na zadatak određivanja

poretka rečenice, a razlikuju se u korištenim korpusima za treniranje i testiranje ili načinu izgradnje rešetke entiteta iz teksta.

U svim eksperimentima provedena je optimizacija SVM hiperparametra *C*. Korištena je optimizacija pretragom po rešetci vrijednosti (engl. *grid search*) uz unakrsnu validaciju (engl. *cross validation*). U ovom slučaju, pošto se radi samo o jednom parametru, rešetka je jednodimenzionalna. Kod optimizacije pretragom po rešetci vrijednosti u ugnježđenim se petljama iterira po mogućim kombinacijama vrijednosti parametara, istrenira se model za trenutnu kombinaciju parametara te se ispita točnost modela. Kod ispitivanja točnosti unakrsnom validacijom, skup za treniranje podjeli se na *n* jednakih dijelova. Za svaki od *n* dijelova, model se istrenira na preostalih *n-1* dijelova, a trenutni dio koristi se za testiranje točnosti modela. Kao ukupna točnost koja se postiže s trenutnom kombinacijom hiperparametara uzima se prosjek *n* točnosti dobivenih u unakrsnoj validaciji.

5.3.1. Osnovni korpusi

Prvi eksperiment jest mjerenje točnosti osnovnog modela rešetke entiteta na dva osnovna korpusa. Model istreniran na korpusu Vjesnik-promet testiran je na korpusu test-Vjesnik-promet, a model istreniran na korpusu Vjesnik-ubojstva testiran je na korpusu test-Vjesnik-ubojstva. Rezultati su prikazani u tablici 5.3.

Tablica 5.3: Točnosti osnovnog modela rešetke entiteta. Model treniran na korpusu Vjesnik-promet testiran je na korpusu test-Vjesnik-promet. Model treniran na korpusu Vjesnik-ubojstva testiran je na korpusu test-Vjesnik-ubojstva.

	promet	ubojstva
točnost	87.75%	93.05%

S ciljem davanja primjera načina računanja točnosti, opisano je kako su dobivene vrijednosti u tablici 5.3. Vrijednost 87.75% dobivena je kao kvocijent 1755/2000. Naime, korpus test-Vjesnik-promet sastoji se od stotinu originalnih tekstova. Svakom tekstu pridruženo je dvadeset različitih permutacija njegovih rečenica. Za svaki par *(original, skup permutacija)* računa se koherentnost originala i permutacija. Ukupno se koherentnost ocjenjuje dvije tisuće i sto puta, a dvije tisuće od tih ocjena koherentnosti su koherentnosti permutacija. Njih tisuću sedamsto pedeset i pet je točno ocjenjeno manje koherentnima od odgovarajućih originala. Odatle dolazi kvocijent 1755/2000.

Rezultat prvog eksperimenta pokazuje značajnu razliku u točnostima dvaju modela. Mogući uzroci te razlike navedeni su poglavlju 5.4. Točnost oba modela usporediva

je ili bolja od rezultata koje postižu Barzilay i Lapata (2008). Njihova dva modela postižu točnosti 87.2% (model *Earthquakes*) te 90.4% (model *Accidents*), no tu se radi o modelima koji koriste alat za automatsko povezivanje koreferentnih imenskih skupova. Njihovi modeli koji koreferenciju računaju grupirajući imenice po identitetu, kao u modelu u ovom radu, postižu točnosti 83.0%, odnosno 89.9%.

Zanimljivo je primjetiti sličnost odnosa točnosti modela *promet* i *ubojstva* naprema odnosu točnosti modela *Earthquakes* i *Accidents*, pogotovo imajući na umu to da je raspodjela duljina tekstova korpusa Vjesnik-promet jednaka raspodjeli korpusa *Earthquakes*, a raspodjela duljina tekstova korpusa Vjesnik-ubojstva jednaka raspodjeli korpusa *Accidents*.[1] Javlja se hipoteza da je utjecaj raspodjele duljine tekstova u korpusima za treniranje i testiranje konzistentan kroz različite jezike i domene.

5.3.2. Podjela po duljini

U drugom eksperimentu proučava se utjecaj varijabilnosti u duljini tekstova na točnost računanja koherentnosti. Korpus Vjesnik-promet, koji sadrži sto dokumenata, podijeljen je u dva korpusa od pedeset dokumenata. Korpus Vjesnik-promet-dugi sadrži pedeset najduljih dokumenata, a korpus Vjesnik-promet-kratki sadrži pedeset najkraćih dokumenata korpusa Vjesnik-promet. Isto je učinjeno za korpus za testiranje test-Vjesnik-promet. Istrenirana su dva modela – na korpusima Vjesnik-promet-dugi i Vjesnik-promet-kratki. Oba modela testirana su zatim na korpusima test-Vjesnik-promet, test-Vjesnik-promet-kratki i test-Vjesnik-promet-dugi. Paralelno tome učinjeno je za korpuse Vjesnik-ubojstva i test-Vjesnik-ubojstva.

Tablica 5.4: Točnosti modela korpusa Vjesnik-promet i podkorpusa dobivenih djeljenjem po duljini. Stupac ćelije odgovara korpusu na kojem je model treniran, a redak ćelije odgovara korpusu na kojem je model testiran.

	Vjesnik-promet-kratki	Vjesnik-promet-dugi
test-Vjesnik-promet	82.35%	80.50%
test-Vjesnik-promet-kratki	82.40%	79.90%
test-Vjesnik-promet-dugi	82.30%	81.10%

U radu Barzilay i Lapata (2008) proveden je eksperiment za proučavanje utjecaja broja tekstova za treniranje na točnost modela. U eksperimentu je model treniran s deset originalnih tekstova, dvadeset, trideset, ... do stotinu tekstova. Kao i ovdje, svakom

[1] Točnije, postoje male razlike u raspodjeli, kao što je objašnjeno u poglavlju 5.2.

Tablica 5.5: Točnosti modela korpusa Vjesnik-ubojstva i podkorpusa dobivenih djeljenjem po duljini. Stupac ćelije odgovara korpusu na kojem je model treniran, a redak ćelije odgovara korpusu na kojem je model testiran.

	Vjesnik-ubojstva-kratki	Vjesnik-ubojstva-dugi
test-Vjesnik-ubojstva	91.85%	90.05%
test-Vjesnik-ubojstva-kratki	92.30%	89.60%
test-Vjesnik-ubojstva-dugi	91.40%	90.50%

originalnom tekstu bilo je pridruženo dvadeset permutacija. Pokazano je da se treniranjem s pedeset originalnih tekstova postiže točnost usporediva s točnošću pri treniranju sa stotinu originalnih tekstova, dok pri treniranju s četrdeset ili manje tekstova točnost opada. Zato je opravdano pretpostaviti da su rezultati u tablicama 5.4 i 5.5 zaista pokazatelji ovisnosti modela o duljinama tekstova. U suprotnom bi se smanjenje točnosti u tablici 5.4 naprema točnosti u tablici 5.3 moglo pripisati smanjenju količine podataka za treniranje.

Rezultati ovog eksperimenta pokazuju smanjenje točnosti pri podjeli tekstova po duljini na kraće i dulje tekstove. Razlika je izraženija za korpus Vjesnik-promet. Razlog za to je vjerojatno distribucija duljine tekstova u korpusu Vjesnik-promet, gdje je raspon duljine tekstova 4–26, što je veće od raspona u korpusu Vjesnik-ubojstva, 6–19. Veći raspon znači i raznovrsniji skup uzoraka prijelaza.

Eksperiment pokazuje, očekivano, da model treniran na dokumentima određene duljine postiže najbolje rezultate na korpusu istih duljina, a najlošije rezultate na korpusu drugačijih duljina. Odnosno, model treniran na kraćim tekstovima ima najmanju točnost pri testiranju na korpusu dugih tekstova, a model treniran na dugim tekstovima ima najmanju točnost pri testiranju na korpusu kraćih tekstova. Zanimljivo je primjetiti da to vrijedi za svaku kombinaciju u obje tablice 5.4 i 5.5.

5.3.3. Kombiniranje tekstova korpusa različitih domena

Treći eksperiment proučava utjecaj domene na točnost modela, odnosno kako se mijenja točnost modela kad se model treniran na jednoj domeni testira na drugoj domeni. Očekivani rezultat jest smanjenje točnosti.

Stvoren je kombinirani korpus Vjesnik-promet-ubojstva koji sadrži pedeset nasumično odabranih tekstova iz korpusa Vjesnik-promet i pedeset nasumično odabranih tekstova iz korpusa Vjesnik-ubojstva. Isto tako je stvoren korpus test-Vjesnik-promet-ubojstva, s time da su iz korpusa test-Vjesnik-promet i test-Vjesnik-ubojstva odabrani

tekstovi jednakih duljina kao tekstovi korpusa Vjesnik-promet-ubojstva.

U sklopu eksperimenta stvorena su tri modela. Prvi model istreniran je na korpusu Vjesnik-promet i testiran na korpusu test-Vjesnik-ubojstva. Drugi model istreniran je na korpusu Vjesnik-ubojstva i testiran na korpusu test-Vjesnik-promet. Treći model istreniran je na korpusu Vjesnik-promet-ubojstva i testiran je na korpusima test-Vjesnik-promet, test-Vjesnik-ubojstva i test-Vjesnik-promet-ubojstva.

Tablica 5.6: Točnosti modela testiranih na korpusima različite domene od korpusa na kojima su trenirani. Stupac ćelije odgovara korpusu na kojem je model treniran, a redak ćelije odgovara korpusu na kojem je model testiran.

	Vjesnik-promet	Vjesnik-ubojstva	Vjesnik-promet-ubojstva
test-Vjesnik-promet	–	86.20%	87.00%
test-Vjesnik-ubojstva	91.95%	–	92.90%
test-Vjesnik-promet-ubojstva	–	–	88.95%

Primjećuje se smanjenje točnosti ukoliko se model treniran na jednoj domeni testira na drugoj domeni. Razlika točnosti nije značajna. Također, razlika u točnostima veća je pri korištenju modela treniranog na tekstovima različitih duljina, a unutar iste domene. Rezultati ovog eksperimenta navode na pitanje koliko su uistinu domenski različiti korpusi Vjesnik-ubojstva i Vjesnik-promet. Za različite domene očekuje se da odgovarajući modeli raspoznaju različite uzorke prijelaza kao koreferentne. Moguće je da su korpusi Vjesnik-ubojstva i Vjesnik-promet visoke međusobne sličnosti unatoč različitim temama tekstova.

5.3.4. Modeli različitih lingvističkih dimenzija

Zadnji eksperimet proučava utjecaj različitih lingvističkih dimenzija na točnost modela. Kao što je opisano u poglavlju 3.3, osnovni model rešetke entiteta sadrži informacije o koreferenciji entiteta, gramatičkoj ulozi entiteta te istaknutosti entiteta u tekstu. No moguće je stvoriti osiromašenu rešetku entiteta koja ne sadrži sve navedene informacije. Neka osnovni model bude označen *COR+SYN+SAL+* jer sadrži informacije koreferencije (engl. *coreference*), sintaksnu informaciju (engl. *syntax*) (gramatička uloga entiteta) te informaciju o istaknutosti entiteta (engl. *salience*). Mogući osiromašeni modeli su: COR+SYN+SAL-, COR+SYN-SAL+, COR+SYN-SAL-, COR-SYN+SAL- i COR-SYN-SAL-. Kombinacija bez koreferencije no sa istaknutosti nije moguća jer ukoliko nema koreferencije, svaki entitet u rešetci entiteta bit će spomenut samo jednom.

Model koji ne sadrži koreferenciju entiteta gradi rešetku entiteta bez grupiranja imenica po identitetu – svaka imenica u tekstu novi je entitet. Model bez sintaksne informacije ne raspoznaje među ulogama entiteta (subjekt, objekt, niti jedno ni drugo) već samo raspoznaje javlja li se entitet u rečenici ili ne. Model bez informacije o istaknutosti ne razlikuje između istaknutih i neistaknutih entiteta, već sve stavlja u istu skupinu.

U skladu s razmišljanjima o lingvističkim dimenzijama rešetke entiteta iz poglavlja 3.3 očekuje se smanjenje točnosti za svaku lingivističku dimenziju koja se oduzme modelu. Naime, model rešetke entiteta vrlo je sažet i sadrži esencijalne informacije, odnosno informacije za koje se teorije koherentnosti bazirane na rešetci entiteta slažu da su esencijalne.

Za osnovne korpuse Vjesnik-promet i Vjesnik-ubojstva istrenirano je svih pet osiromašenih modela koji su potom testirani na odgovarajućem korpusu za testiranje.

Tablica 5.7: Točnosti modela za različite osiromašene verzije modela rešetke entiteta. Pritom "COR" znači koreferencija, "SYN" – sintaksa, "SAL" – istaknutost. Primjerice "COR-" označava da model ne sadrži informacije o koreferenciji entiteta.

	COR+SYN+SAL-	COR+SYN-SAL+	COR+SYN-SAL-	COR-SYN+SAL-	COR-SYN-SAL-
promet	88.50%	88.50%	90.15%	89.90%	88.80%
ubojstva	91.95%	93.45%	92.15%	92.75%	90.35%

Rezultati ovog eksperimenta neočekivani su. Za model *promet* svaki osiromašeni model postiže veću točnost od osnovnog modela. Svi modeli postižu veću točnost ukoliko se isključi informacija o istaknutosti entiteta. Značenje sintaksne informacije nije jednoznačno: modeli COR+SYN-SAL+ i COR+SYN-SAL- postiže veću točnost odgovarajućih modela bez sintaksne informacije, no model COR-SYN+SAL- postiže veću točnost od modela COR+SYN-SAL-. Uloga koreferencije nije jednoznačna kao ni uloga sintekse. Koreferencija ne podiže uvijek točnost: model COR-SYN+SAL- postiže veću točnost od modela COR+SYN-SAL-.

Za model *ubojstva* vrijede ista opažanja o nejednoznačnim ulogama koreferencije i sintakse kao za model *promet*, no ovdje niti uloga istaknutosti nije jednoznačna. Također, samo model COR+SYN-SAL+ postiže veću točnost od osnovnog modela.

5.4. Analiza rezultata

Eksperimenti iz poglavlja 5.3 polučuju nekoliko zanimljivih rezultata. Prvi eksperiment (poglavlje 5.3.1) pokazao je da model opisan u ovom radu postiže veću točnost

od modela Barzilay i Lapata (2008) kad njihov model koristi punu koreferenciju imenskih skupova ili koreferenciju pomoću grupiranja jednakih imenica. Precizna usporedba nije moguća bez analize korpusa jer modeli nisu trenirani i ispitivani na istim korpusima, ili na korpusima koji imaju mnogo zajedničkih svojstava.

Drugi eksperiment (poglavlje 5.3.2) pokazao je da unutar jedne domene tekstova postoje različiti uzorci prijelaza entiteta u duljim i kraćim tekstovima. Model rešetke entiteta jest model lokalne koherentnosti i nije predviđen za dugačke tekstove (iako nije formalno definirano što se smatra dugačkim tekstom), što znači da će skup korpusa na kojima se model može koristiti biti odozgor ograničen duljinom teksta. Implikacija drugog eksperimenta jest da je pri izradi modela za određenu domenu potrebno eksperimentirati s raspodjelama duljina tekstova u skupu za treniranje kako bi se dobili najbolji rezultati.

U drugom eksperimentu, modeli trenirani na podskupovima korpusa Vjesnik-promet postižu značajno niže rezultate od modela podskupova korpusa Vjesnik-ubojstva. Kao moguće objašnjenje dana je razlika u raspodjelama duljina tekstova. Usporedbom sa rezultatima prvog eksperimenta, dolazim do pretpostavke da na razliku u točnosti modela u prvom eksperimentu utječu jedino razlike u raspodjeli duljina. Rezultati trećeg eksperimenta mogu podržati takvo razmišljanje. U njemu je pokazano da primjena modela treniranog u domeni *promet* postiže tek malo slabije rezultate kad se primjeni na drugu domenu, *ubojstva*, te obratno. Vjerojatno to ne znači da su uzorci prijelaza entiteta neovisni o domeni, nego da su korpusi Vjesnik-ubojstva i Vjesnik-promet previše slični i da se između njih ne može provesti valjan eksperiment o utjecaju kombiniranja domena na točnost modela. To donekle narušava smisao korištenja tih dvaju korpusa u radu, jer smisao je bio koristiti korpuse s različitim uzorcima prijelaza entiteta. No, zbog razlike u raspodjelama duljina između tih dvaju korpusa, pokazan je utjecaj raspodjele duljina na točnost modela.

Rezultati zadnjeg eksperimenta iznenađujući su i pokazuju da na korpusima Vjesnik-ubojstva i Vjesnik-promet nije isplativo uvoditi lingvističko znanje osnovnog modela rešetke entiteta. Siromašan model COR-SYN-SAL- postiže rezultate usporedive s rezultatima punog modela COR+SYN+SAL+.

Moguće objašnjenje rezultata zadnjeg eksperimenta jest da je korpus takav da su tekstovi naprosto previše slični. Oduzimanje ekspresivnosti vektora značajki zbog toga ima manji utjecaj na točnost od smanjenja dimenzionalnosti vektora koje se pritom događa. Zbog toga je moguće naučiti modele koji postižu visoku točnost na danom skupu tekstova za učenje. Traženje dokaza za ovu hipotezu zahtjeva korpusnu analizu, a to je izvan opsega ovog rada.

5.4.1. Najčešći prijelazi

U analizi pogrešaka modela zanimljivo je analizirati slučajeve za koje model krivo radi, s ciljem stjecanja uvida u moguća poboljšanja, ograničenja i općenito rad modela. To je inherentno teško učiniti za model rešetke entiteta, jer bi bilo potrebno analizirati stotine permutiranih tekstova i odgovarajućih originala.[2] Ta analiza zahtjevala bi i dublje proučavanje veze uzoraka prijelaza entiteta i koherentnosti tekstova, iz perspektive teorija koherentnosti. Sve to je izvan opsega Završnog rada te je unutar ove analize rezultata odabran jednostavniji pristup.

U tablicama 5.8 i 5.9 navedeno je dvadeset najčešćih prijelaza entiteta u rešetkama entiteta za originalne tekstove, permutacije tekstova i permutacije ocjenjene koherentnijima od pripadnih originala, za korpuse test-Vjesnik-ubojstva i test-Vjesnik-promet. Cilj je vidjeti koji su prijelazi svojstveni tekstovima koje model ocjenjuje koherentnima, a koji tekstovima koje ocjenjuje nekoherentnima.

Očekivano, dva daleko najčešća prijelaza za sve skupine tekstova su [–,–] te [–,–,–]. Rešetke entiteta dakle uglavnom sadrže velik broj gotovo praznih stupaca, kao što je predviđeno u poglavlju 3.2.

Usporedbom stupaca u obje tablice vidi se veća sličnost 1. i 3. stupca nego 1. i 2. ili 2. i 3. stupca. U tablici 5.9 jednako je prvih sedam redaka 1. i 3. stupca, dok je 2. stupac različit već u 3. retku. U tablici 5.8 jednako je prvih devet redaka 1. i 3. stupca, a 2. stupac otklanja se u 5. retku. To je logično, s obzirom da model ocjenjuje tekstova s prijelazima u ta dva stupca koherentnima.

Zanimljivo je primijetiti razliku između korpusa Vjesnik-ubojstva i Vjesnik-promet, realiziranu time što su za korpus Vjesnik-promet češći prijelazi koji sadrže subjekt, nego prijelazi koji sadrže objekt, dok obratno vrijedi za korpus Vjesnik-ubojstva. Dakle, primjedba da su ta dva korpusa vrlo slična ipak nije posve točna.

[2] Preciznije, bilo bi potrebno proučiti sve permutirane dokumente ocjenjene koherentnijima od pripadnih originala te pripadne originale.

Tablica 5.8: Najčešćih dvadeset prijelaza u korpusu test-Vjesnik-ubojstva. Stupac "originali" sadrži prijelaze iz originalnih dokumenata. Stupac "sve permutacije" odnosi se na sve permutacije originalnih dokumenata (njih dvije tisuće). Stupac "koherentne permutacije" sadrži prijelaze permutacija dokumenata koje je model ocjenio koherentnijima od njihovih originala.

	originali	sve permutacije	koherentne permutacije
1	–,–	–,–	–,–
2	–,–,–	–,–,–	–,–,–
3	–,X	–,X	–,X
4	X,–	X,–	X,–
5	–,X,–	–,–,X	–,X,–
6	X,–,–	–X,–	X,–,–
7	–,–,X	X,–,–	–,–,X
8	–,O	O,–	–,O
9	O,–	–,O	O,–
10	–,–,O	–,–,O	–,O,–
11	–,O,–	–,O,–	–,–,O
12	O,–,–	O,–,–	O,–,–
13	–,S	S,–	–,S
14	S,–	–,S	–,–,S
15	–,–,S	–,S,–	S,–
16	–,S,–	S,–,–	–,S,–
17	S,–,–	–,–,S	S,–,–
18	X,X	X,X	X,X
19	X,X,–	X,–,X	X,X,–
20	–,X,X	X,X,–	–,X,X

Tablica 5.9: Najčešćih dvadeset prijelaza u korpusu test-Vjesnik-promet. Stupci "originali", "sve permutacije" i "koherentne permutacije" imaju isto značenje kao u tablici 5.8.

	originali	sve permutacije	koherentne permutacije
1	–,–	–,–	–,–
2	–,–,–	–,–,–	–,–,–
3	–,X	X,–	–,X
4	–,X,–	–,X	–,X,–
5	X,–	X,–,–	X,–
6	–,–,X	–,X,–	–,–,X
7	X,–,–	–,–,X	X,–,–
8	–,S	S,–	–,S
9	S,–	–,S	S,–
10	–,–,S	S,–	–,–,S
11	–,S,–	–,S,–	–,S,–
12	–,O	–,–,S	S,–,–
13	–,O,–	O,–	–,O
14	–,–,O	–,O	O,–
15	O,–	–,O,–	–,O,–
16	S,–,–	–,X,X	–,–,O
17	O,–,–	O,–,–	O,–,–
18	X,X	X,X	X,X
19	–,X,X	X,X,–	–,X,X
20	X,X,–	–,X,X	X,X,–

6. Zaključak

Sustavi za strojno generiranje teksta često koriste alate kojima rangiraju potencijalne rezultantne tekstove s obzirom na određeno bitno svojstvo teksta te potom za krajnji rezultat odabiru tekst s najvišim rangom. Jedno od takvih svojstava teksta jest koherentnost, značajka teksta ključna za njegovu razumljivost.

U ovom radu prezentiran je model lokalne koherentnosti tekstova na hrvatskom jeziku temeljen na rešetci entiteta. Model je napravljen da funkcionira posve automatski, učeći vjerojatnosti prijelaza entiteta svojstvene koherentnim tekstovima direktno iz tekstova korpusa, bez potrebe za ručnim označavanjem podataka za bilo koji dio modela, zbog čega je model moguće vrlo jednostavno i brzo prenijeti na nove domene.

U radu je provedno nekoliko eksperimenata – zadatak određivanja poretka rečenica s različitim postavkama. Model postiže visoku točnost: 87.75% te 93.05% na dva korpusa, što je rezultat usporediv s točnošću koju postiže model u originalnom radu za engleski jezik: 87.2%, 90.4% (Barzilay i Lapata, 2008). Rezultati eksperimenata ukazuju i na važnost utjecaja distribucije duljine tekstova u korpusu za treniranje na točnost modela unutar određene domene.

Nekoliko je mogućih nastavka na ovaj model. Jedan smjer je ukomponirati ovaj model u sustav za strojno generiranje teksta na hrvatskom jeziku. Druga mogućnost je inkomponirati u model nadogradnje na osnovni model rešetke entiteta s ciljem povećanja točnosti – primjerice uvesti stupnjevanje nekoherentnosti permutacija, umjesto da se sve permutacije smatraju jednako nekoherentnima. Također bi bilo zanimljivo istrenirati i testirati model na većem broju korpusa iz različitih domena, s ciljem bolje analize utjecaja domene na model.

LITERATURA

Željko Agić i Danijela Merkler. Three syntactic formalisms for data-driven dependency parsing of croatian. *Text, Speech and Dialogue. Lecture Notes in Computer Science*, stranice 560–567, 2013.

Željko Agić, Nikola Ljubešić, i Danijela Merkler. Lemmatization and morphosyntactic tagging of croatian and serbian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, stranice 48–57, 2013.

Regina Barzilay i Mirella Lapata. Modeling local coherence: An entity-based approach. *Proceedings of the 43rd Annual Meeting of the ACL*, stranice 141–148, 2005.

Regina Barzilay i Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

Regina Barzilay i Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. *HLT-NAACL 2004: Proceedings of the Main Conference*, stranice 113–120, 2004.

Walter Foltz, Peter W.; Kintsch i Thomas K. Landauer. The measurement of local coherence with latent semantic analysis. *Discourse Processes*, 25(23):285–307, 1998.

Scott Grosz, Barbara J; Weinstein i Aravind K Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995.

Thorsten Joachims. Training linear svms in linear time. U *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, stranice 217–226. ACM, 2006.

Ziheng Lin, Hwee Tou Ng, i Min-Yen Kan. Automatically evaluating text coherence using discourse relations. U *Proceedings of the 49th Annual Meeting of the Asso-*

ciation for Computational Linguistics: Human Language Technologies-Volume 1, stranice 997–1006. Association for Computational Linguistics, 2011.

Nikola Ljubešić i Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. U *Text, Speech and Dialogue*, stranice 395–402. Springer, 2011.

William C. Mann. Intro to rst. `http://www.sfu.ca/rst/01intro/intro.html`. Accessed: 2015-06-05.

William C. Mann i Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. 1988.

Daniel Marcu. Distinguishing between coherent and incoherent texts. U *The Proceedings of the Student Conference on Computational Linguistics in Montreal*, stranice 136–143, 1996.

Hwanjo Yu, Youngdae Kim, i Seungwon Hwang. Rv-svm: An efficient method for learning ranking svm. U *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings*, svezak 5476, stranica 426. Springer, 2009.

MODEL KOHERENTNOSTI TEKSTOVA NA HRVATSKOM JEZIKU TEMELJEN NA ANALIZI ENTITETA

Sažetak

Koherentnost je značajka prirodnog testka koja direktno utječe na razumljivost teksta, a opisuje povezanost djelova teksta u cjelinu. U radu je prezentiran model lokalne koherentnosti temeljen na rešetci entiteta za tekstove na hrvatskom jeziku. Algoritam iz tekstova vadi entitete i njihove gramatičke uloge te apstrahira tekst u rešetku prijelaza uloga entiteta te na sintetički označenom korpusu uči raspodjelu prijelaza karakterističnu za koherentne tekstove. Rezultantni model rangira ulazne tekstove po koherentnosti. Model postiže visoku točnost na zadatku određivanja poretka rečenice, 87%-93% na kratkim novinskim člancima.

Ključne riječi: obrada prirodnog jezika, hrvatski jezik, diskurs, rešetka entiteta, lokalna koherentnost, entitet

Entity-Based Coherence Model for Croatian Texts

Abstract

Coherence is a feature of natural texts that directly influences its understandability and corresponds to the level of connectedness of parts of texts into a logical whole. This thesis presents an entity grid based model of local coherence for Croatian texts. The algorithm extracts entites and their grammatical roles from texts and abstracts the text into a grid of entity role transitions. The model learns distributions of transitions characteristic for coherent texts on synthetically annotated corpora. The resulting model rangs input texts by coherence. A high performance is achieved by the model on the sentence ordering task, 87%-93% on short newspaper articles.

Keywords: natural language processing, Croatian language, discourse, entity grid, local coherence, entity-based