Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 605

Razrješavanje koreferencije u tekstovima na hrvatskome jeziku

Matija Hanževački

Zagreb, lipanj 2013.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 11. ožujka 2013.

DIPLOMSKI ZADATAK br. 605

Pristupnik: **Matija Hanževački**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Razrješavanje koreferencije u tekstovima na hrvatskome jeziku**

Opis zadatka:

Razrješavanje koreferencije postupak je kojim se utvrđuje koji se izrazi u tekstu dokumenta odnose na isti izvanjezični entitet. Koreferentni izrazi mogu biti vlastita imena, imeničke fraze ili zamjenice. Razrješavanje koreferencije važan je zadatak u okviru obrade prirodnog jezika te nužan preduvjet za mnoge zadatke ekstrakcije informacija. Radi se o izrazito semantičkom problemu koji je težak kako za označavanje podataka, tako i za automatizirano rješavanje i vrednovanje.
U okviru diplomskog rada potrebno je proučiti postupke i sustave za razrješavanje koreferencija u tekstu. Razraditi postupak za otkrivanje referentnih spominjanja i razrješavanje koreferencije u tekstovima na hrvatskome jeziku, uzimajući u obzir nedostatak dostupnih jezičnotehnoloških alata za hrvatski jezik. Postupak se treba temeljiti na metodama strojnog učenja te kombinirati klasifikaciju parova spominjanja i grupiranje referentnih spominjanja. Razviti programsku implementaciju postupka i primijeniti ga na označenom novinskom korpusu tekstova na hrvatskome jeziku. Provesti eksperimentalno vrednovanje točnosti ekstrakcije, analizu značajki te detaljnu analizu pogrešaka. Radu priložiti izvorni programski kod, programsku dokumentaciju i označene skupove podataka.

Zadatak uručen pristupniku: 15. ožujka 2013.
Rok za predaju rada: 28. lipnja 2013.

Mentor:

Doc.dr.sc. Jan Šnajder

Djelovođa:

Doc.dr.sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof.dr.sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Obrada prirodnog jezika (engl. *natural language processing*) je interdisciplinarno područje koje se smatra dijelom računarske znanosti, umjetne inteligencije i lingvistike. Ona se bavi interakcijom između računala i čovjeka putem prirodnog jezika, a neki od najpoznatijih teških zadataka u ovom području su strojno prevođenje (engl. *machine translation*), automatsko odgovaranje na pitanja (engl. *automatic question answering*), automatsko sažimanje teksta (engl. *automatic text summarization*), i razumijevanje prirodnog jezika (engl. *natural language understanding*).

Područje koje se bavi dohvatom strukturiranih informacija iz nestrukturiranih ili polustrukturiranih tekstnih podataka je ekstrakcija informacija (engl. *information extraction*). Razrješavanje koreferencije (engl. *entity coreference resolution*) među imeničkim frazama je problem koji spada upravo u skup zadataka koje obuhvaća ekstrakcija informacija i podrazumijeva otkrivanje spominjanja izvanjezičnih entiteta u tekstu i njihovo grupiranje.

Entitetom (engl. *entity*) se smatra bilo što što postoji samo po sebi, odnosno nešto čemu se može pridijeliti identitet. To su primjerice osobe, životinje, stvari, pojave, lokacije, ali i pravni i poslovni pojmovi poput trgovačkih društava i sl. Spominjanja (engl. *mentions*) su pak izrazi u tekstu, odnosno najčešće imeničke fraze (engl. *noun phrases, NP*), koje se koriste kao referenca na neki entitet u stvarnom svijetu. Par spominjanja u tekstu se smatra koreferentnim ako se oba spominjanja odnosne na identični vanjezični entitet. Grupe spominjanja koje se odnose na isti entitet predstavljaju taj entitet u samom tekstu i cilj je ovog zadatka otkriti takva spominjanja i grupirati ih u grupe. Primjer koreferentnih spominjanja dan je u nastavku, gdje su spominjanja jednog entiteta označena kurzivom, a drugog podebljanjem.

„*Hrvatski predsjednik Ivo Josipović* naglasio je važnost uključenja mladih u ovaj projekt. *Predsjednik* je ujedno naglasio potrebu za pokretanjem sličnih projekata. **Predsjednik udruge Marko Marić** ovom *mu* je prilikom uručio plaketu, na čemu **mu** je *predsjednik* zahvalio i rekao da *on* tako nešto nije očekivao.“

Rješavanje ovog problema važan je korak prema rješavanju ranije navedenih problema u obradi prirodnog jezika s obzirom da je vrlo korisno za razumijevanje otkriti entitete, odnosno njihova spominjanja u tekstu. Nakon toga se iz teksta mogu ekstrahirati relacije među tim entitetima, što je daljnji korak prema računalnom razumijevanju tekstova na prirodnom jeziku.

Ovaj zadatak u engleskom jeziku pretpostavlja korištenje naprednih jezičnih alata poput sintaktičkih parsera (engl. *syntax parser*), sustava za prepoznavanje imenovanih entiteta (engl. *named entity recognition*), ali i označivača vrsta riječi (engl. *part of speech tagger*) i lematizatora (engl. *lemmatizer*). Nedostatak ovih alata je do nedavno uvelike otežavao izradu kvalitetnog sustava za razrješavanje koreferencije u tekstovima na hrvatskome jeziku. U međuvremenu su navedeni alati postali dostupni te su iskorišteni u oblikovanju ovog sustava. Koliko je poznato autoru, jedini radovi koji su se bavili razrješavanjem koreferencije u hrvatskome jeziku su Kmetovic (2011); Hranj (2011).

U ovom je radu opisan sustav za razrješavanje koreferencije u tekstovima na hrvatskome jeziku. Sustav se temelji na nadziranom statističkom strojnom učenju pri čemu je korišten vrlo popularan pristup, tzv. model parova spominjanja (engl. *mention-pair model*), koji su prvi put predložili Aone i Bennett (1995) i McCarthy i Lehnert (1995). Kandidati za spominjanja su iz teksta izvučeni metodom temeljenom na ručno oblikovanim pravilima. Kao skup podataka za učenje iskorišten je prethodno označeni korpus novinskih tekstova na hrvatskome jeziku. Za vrednovanje postupka iskorišten je službeni evaluator[1] konferencije *CoNLL* 2011 (engl. *Conference on Computational Natural Language Learning 2011*).

U sljedećem je poglavlju dan pregled područja i srodnih radova. Nakon toga je u trećem poglavlju detaljnije opisan problem razrješavanja koreferencije. Opis korištenog skupa podataka dan je u četvrtom poglavlju. U petom su poglavlju navedene pojedinosti o razvijenom sustavu za razrješavanje koreferencije, dok je u poglavlju nakon toga opisano programsko ostvarenje sustava. Zatim su u idućem poglavlju dani rezultati vrednovanja uspješnosti i analize pogrešaka. U osmom poglavlju izložen je zaključak zajedno s planovima za budući rad.

[1] `http://conll.cemantix.org/2011/software.html`

2. Pregled područja

Nadzirane metode za razrješavanje koreferencije pojavile su se prije gotovo dvadeset godina, a prvih petnaest godina detaljno je obrađeno u (Ng, 2010). U ovom je poglavlju stavljen naglasak upravo na razne metode temeljene na nadziranom strojnom učenju zajedno s naglaskom na trenutno najbolje sustave na ovom području. Također je spomenuto nekoliko novijih radova koji su se pozabavili drugačijim korpusima u odnosu na čiste i jasne novinske tekstove, odnosno korisnički generiranim sadržajem.

Trenutno najboljim sustavom za razrješavanje koreferencije (engl. *state-of-the-art*) smatra se Stanfordov sustav s konferencije *CoNLL* 2011, od Lee et al. (2011). To je sustav u potpunosti temeljen na pravilima, a zasniva se na ranijem sustavu od Raghunathan et al. (2010). Ovaj je sustav pobijedio na konferenciji *CoNLL* 2011, s rezultatom od $F1 = 58.3\%$ koristeći automatski ekstrahirana spominjanja, odnosno rezultatom od $F1 = 61.4\%$ koristeći ručno označena spominjanja. Opis načina vrednovanja dan je u poglavlju o vrednovanju uspješnosti, s obzirom da se jednak način vrednovanja koristi i u ovom radu.

Sustav se temelji na višeprolaznom „situ“ koje u svakom prolazu implementira filtriranje sa sve manjom i manjom preciznosti. Sustav je podijeljen u tri glavne faze: ektrakciju spominjanja, razrješavanje koreferencije i završnu obradu. Veći se odziv preferira pri ekstrakciji spominjanja, dok se preciznost favorizira u fazi razrješavanja koreferencije. U fazi završne obrade odbacuju se entiteti sa samo jednim spominjanjem (engl. *singleton*).

2.1. Dostupni korpusi

Korpusi su skupovi tekstova s najčešće ručno dodanim korisnim metapodacima. Ti metapodaci predstavljaju razne informacije o tekstu i njegovim dijelovima. Poput: oznaka vrsta riječi, sintaktičkih informacija, imenovanih entiteta, koreferencija i sl. Korpusi služe kao skupovi podataka za učenje klasifikatora, ali i za evaluaciju. Njihovo oblikovanje zahtijeva velike količine resursa, ovisno o veličini željenog korpusa, ali i

težini zadatka.

Neki od najviše korištenih korpusa s označenim koreferencijama su i *MUC* (engl. *Message Understanding Conference*) korpusi. *MUC*6 i *MUC*7 korpusi, svaki od po 60 dokumenata novinskih tekstova, svojevremeno su bili vrlo popularni u izradi sustava za razrješavanje koreferencije (Ng, 2010). Korpusi s konferencija *ACE* (engl. *Automatic Content Extraction*) su otprilike jednako popularni. Raniji *ACE* korpusi, poput *ACE*-2, se sastoje od isključivo engleskih novinskih i televizijskih članaka, dok su kasnije verzije, poput *ACE* 2005, uključile i tekstove na kineskom i arapskom jeziku, i to iz raznih izvora poput radijskih razgovora, blogova, useneta i sl. (Ng, 2010). *ACE* 2005 korpus sastoji se od ukupno 600 dokumenata.

Jedan od najbitnijih novih korpusa za engleski jezik je *Ontonotes* korpus, korišten na konferenciji *CoNLL* 2011 i opisan u (Pradhan et al., 2011). Veličine otprilike 1.3M pojavnica, s označenim entitetima i događajima, bez ograničenja spominjanja na imeničke fraze. *Ontonotes* također sadrži i tekstove na kineskom i arapskom jeziku.

Koreferencija je označena i u nekim bankama stabala. Tuebingen, njemačka banka stabala od 27 tisuća riječi, sastoji se od njemačkih novinskih tekstova i dostupna je od 2004. godine. Češka praška banka stabala verzije 2 sastoji se od ukupno 755 dokumenata i opisana je u (Nedoluzhko et al., 2009). NAIST korpus sastoji se od 287 japanskih novinskih članaka, a tu su zatim i CESS-ECE, opisan u (Recasens et al., 2007) te AnCora korpus iz 2009. koji se sastoje od španjolskih i katalonskih tekstova. Bugarski korpus, veličine 50 tisuća riječi također je dostupan javnosti. Poljski korpus, opisan u (Broda et al., 2012), sastoji se od 1458 dokumenata i otprilike 400 tisuća pojavnica. Također postoji i novi baskijski korpus opisan u (Soraluze et al.) od ukupno 662 spominjanja.

2.2. Otkrivanje spominjanja

Problem razrješavanja koreferencije može se ugrubo podijeliti na dva glavna podzadatka: otkrivanje i ekstrakciju spominjanja te grupiranje spominjanja po entitetima. Bez obzira što potpuni zadatak razrješavanja koreferencije zahtijeva rješavanje oba ova podzadatka, pojavili su se radovi koji su se koncentrirali samo na problem ekstrakcije spominjanja iz teksta.

Kummerfeld et al. (2011) reducirali su sustav koji su razvili Haghighi i Klein (2010) kako bi poboljšali otkrivanje spominjanja. To su ostvarili dodavanjem raznih filtara temeljenih na ručno oblikovanim pravilima, poput uklanjanja generičkih riječi, određenih oblika brojeva i određenih vrsta riječi. Soraluze et al. predstavljaju

sustav za otkrivanje spominjanja na baskijskom, temeljen na ručno oblikovanim pravilima koja su potom prevedena u strojeve s konačnim brojem stanja. Florian et al. (2010) predstavljaju sustav za poboljšanje otkrivanja spominjanja na dokumentima s ne nužno čistim i pravilnim engleskim jezikom. Automatski detektiraju na kojoj je jezičnoj razini tekst i po tome odlučuju hoće li ga obraditi klasičnom metodom ili će ga preuzeti metoda koja koristi visoko semantičke značajke s klasifikatorom najveće entropije (engl. *maximum entropy classifier*).

2.3. Nadzirani modeli razrješavanja koreferencije

Postoje tri glavne vrste modela pristupa razrješavanje koreferencije pomoću nadziranog strojnog učenja (engl. *supervised machine learning*) prema Ng (2010). To su: model parova spominjanja (engl. *mention-pair model*), model entiteta i spominjanja (engl. *entity-mention model*) i rangirajući model (engl. *ranking model*). Svaki od spomenutih modela detaljnije je opisan u nastavku.

Neki od najčešće korištenih algoritama za učenje ovih modela prema Ng (2010) su algoritmi stabala odluke (engl. *Decision Trees*) (primjerice C5), algoritmi za učenje pravila (npr. RIPPER), zatim modeli najveće entropije (engl. *maximum entropy models*), neuronske mreže (engl. *neural networks*) i strojevi s potpornim vektorima (engl. *support vector machines*).

2.3.1. Model parova spominjanja

Model parova spominjanja je najpopularniji model nadziranog razrješavanja koreferencije koji se temelji na binarnoj klasifikaciji parova spominjanja. Model promatra sve parove spominjanja neovisno jedan od drugome, u dokumentu i donosi odluku o njihovoj mogućoj koreferenciji.

Neke od glavnih slabosti ovoga modela su:

- *Nedostatak tranzitivnosti relacije koreferencije* – svaki se par spominjanja promatra neovisno, što zahtijeva odvojeno grupiranje parova spominjanja;
- *Nesrazmjerna razdioba klasa* – promatraju se svi parovi spominjanja u tekstu pa postoji nesrazmjerno velik broj parova koji nisu koreferentni u odnosu na one koji to jesu;
- *Neovisno promatranje parova spominjanja* – ne postoji nikakav način određivanja najvjerojatnijeg antecedenta, a ni trenutna odluka o klasifikaciji ne ovisi

o prethodnim odlukama, što predstavlja problem kod kasnije faze grupiranja;

– *Nedostatak konteksta* – informacije ekstrahirane iz samo dvije imeničke fraze vrlo vjerojatno nisu dovoljne za pouzdanu klasifikaciju.

Iz razloga što je nerealno očekivati da se u dokumentu promatraju apsolutno svi parovi spominjanja (što je reda veličine n^2 parova), u implementaciji ovog modela se primjenjuju razne heuristike za odabir primjera za učenje. Ove su heuristike uglavnom usmjerene u smanjivanje nesrazmjera među klasama, gdje inače ima puno više negativnih od pozitivnih primjera. Jedna od najpopularnijih strategija jest strategija koju su predložili Soon et al. (2001) gdje se pozitivna instanca određuje između svaka dva susjedna spominjanja nekog entiteta, a negativni primjeri se generiraju tako da se stvore parovi trenutnog spominjanja i svih kandidata između trenutnog spominjanja i njegovog stvarnog antecedenta.

Ng i Cardie (2002) predlažu još jedan dodatni uvjet koji definira da ako je trenutno spominjanje zamjeničko, onda je nužno potrebno da je njegov antecedent imenski. Strube et al. (2002) predlažu filtriranje parova spominjanja prema slaganju po rodu i broju, odnosno očekuje se da oba spominjanja imaju međusobno jednak rod i broj. Pristup koji uključuje sve gore navedene heuristike smatra se trenutno najboljim postupkom za odabir primjera za učenje.

Po završetku klasifikacije parova potrebno je obaviti grupiranje spominjanja u entitete. Pokazalo se da jednostavna pretpostavka tranzitivnosti ne daje dovoljno dobre rezultate (Ng, 2010) pa se umjesto toga koriste razni algoritmi za grupiranje. Neki od najčešće korištenih su *algoritam najbliži prvi* i *algoritam najbolji prvi*. Neki od glavnih nedostataka ovih metoda su njihova pohlepnost[1] (engl. *greediness*) pa su umjesto toga predložene razne složenije strategije grupiranja, poput *korelacijskog grupiranja*, *algoritama podjele grafova* i sl. Bez obzira na zamjetan broj predloženih metoda grupiranja, još ne postoje kvalitetne usporedbe performansi navedenih metoda.

2.3.2. Model entiteta i spominjanja

Jedan od glavnih problema modela parova spominjanja je činjenica da trenutna odluka o klasifikaciji para spominjanja ne ovisi o prethodnim odlukama. Model entiteta i spominjanja rješava taj problem oblikujući model koji donosi odluku o klasifikaciji spominjanja s jedne strane i prethodno (djelomično) oblikovanih grupa spominjanja,

[1] Pohlepni algoritmi su oni koji povlače samo lokalno optimalne poteze, u nadi da će pronaći globalni optimum.

odnosno entiteta, s druge strane. Ovakav model implicira korištenje značajki na razini entiteta, što povećava ekspresivnost u odnosu na model parova spominjanja.

Zanimljivo je primjetiti da prema Ng (2010) sustavi temeljeni na ovom modelu nisu pokazali značajno bolje (ili su čak pokazali lošije) rezultate od modela parova spominjanja, usprkos teoretskim prednostima ovoga modela.

2.3.3. Rangirajući model

Rangirajući se modeli zasnivaju na određivanju najvjerojatnijeg antecedenta za danu imeničku frazu. Na taj se način istovremeno razmatraju svi kandidati za antecedenta i modelira se natjecanje među njima. Neki modeli navedeni u (Ng, 2010) implementiraju rangiranje prethodno oblikovanih grupa, po uzoru na model entiteta i spominjanja. Na taj se način rješavaju neke od glavnih slabosti modela parova spominjanja, poput neovisnog promatranja parova spominjanja prilikom klasifikacije i nedostatka tranzitivnosti.

2.3.4. Značajke

Prema Ng (2010); Soon et al. (2001); Ng i Cardie (2002), najčešće korišteni tipovi značajki u nadziranim modelima za razrješavanje koreferencije su:

- *Podudaranje nizova znakova* – Osim klasičnog potpunog ili djelomičnog podudaranja nizova znakova, koriste se još i podudaranje glavnih riječi imeničkih fraza, zatim najmanji broj promjena potrebnih za transformaciju jednog niza u drugi (engl. *minimum edit distance, MED*), najdulji zajednički podniz i sl.;
- *Sintaktičke značajke* – Razne značajke koje se ektrahiraju iz strukturnih (engl. *constituency parser*) i ovisnosnih (engl. *dependency parser*) sintaktičkih analizatora. Primjerice, ekstrahiraju se putevi po sintaksnom stablu između imeničkih fraza ili se uvode mjere sličnosti sintaksnih stabala spominjanja;
- *Gramatičke značajke* – Kodiraju gramatička svojstva imeničkih fraza, poput tipa imeničke fraze ili roda i broja, gdje se očekuje strogo slaganje između dvije fraze u tom pogledu;
- *Semantičke značajke* – Neke od najpopularnijih značajki iz ove grupe su tzv. odabirne preferencije (engl. *selectional preference*) gdje je uz zamjenicu dan i njezin odgovarajući glagol (engl. *governing verb*) za koji se onda traži imenski antecedent s istim glagolom. Također se ekstrahiraju semantičke značajke iz raznih skupova lokacija (engl. *gazzetteer*) ili skupova sinonima (primjerice

WordNet[2]) ili Wikipedije, pomoću kojih se računa sličnost između dviju fraza. Tu su još i semantičke klase poput *osobe, tvrtke, lokacije, objekta* koje se ekstrahiraju sustavima za ekstrakciju imenovanih entiteta (engl. *named entity recognition, NER*) i onda primjenjuju između ostalog u ekstrakciji spominjanja;

– *Leksičko-sintaktički uzorci* – Ovakvi uzorci modeliraju semantičke odnose među imeničkim frazama. Primjerice, ako imenička fraza odgovara nekom ručno rađenom uzorku, vrlo je vjerojatno da i druga fraza koja odgovara tom istom uzorku jest u koreferenciji s prvom. Što se češće taj uzorak pojavljuje, veća je vjerojatnost koreferencije;

– *Diskursne značajke* – Tipično uključuju udaljenost između dvije imeničke fraze u tekstu ili indikaciju da je neka fraza subjekt i sl.;

– *Ostale* – Ostale značajke uključuju izlaze sustava za otkrivanje spominjanja ili razrješavanje koreferencije, temeljenih na pravilima, koji se onda mogu koristiti za polu-nadzirano učenje (engl. *bootstrapping*). Tu su naravno i oznake vrsta riječi (engl. *part of speech, POS, tags*).

2.4. Razrješavanje koreferencije u korisnički generiranom sadržaju

Svega se nekoliko radova do sada bavilo razrješavanjem koreferencije u korisnički generiranom sadržaju. Ovakav pristup problemu je zanimljiv jer se na internetu svakim danom generiraju sve veće količine korisničkog sadržaja, koji nije pisan na čistom standardnom jeziku, već je pun šuma i raznih gramatičkih i pravopisnih pogrešaka i kolokvijalizama. Takav jezik predstavlja velik problem postojećim sustavima za razrješavanje koreferencije koji su oblikovani i trenirani na strukturiranim novinskim korpusima. Zbog toga se pojavila potreba za oblikovanjem robusnih sustava koji se mogu nositi s takvim jezikom blogova, društvenih mreža ili govornim jezikom.

Hendrickx i Hoste (2009) opisuju sustav za razrješavanje koreferencije na tekstovima s blogova i komentiranih novinskih članaka na nizozemskom jeziku. Takvi su tekstovi primjer vrlo nestrukturiranih tekstova s puno šuma koji otežavaju ekstrakciju pouzdanih informacija o semantici promatranog teksta. Koriste model parova spominjanja i pokazuju drastičan pad performansi sustava po mjeri *MUC* pri promjeni korpusa. Preciznije, s novinskih članaka ($F1 = 53.6\%$), na blogove ($F1 = 25.7\%$), a

[2] `http://wordnet.princeton.edu/`

potom i na komentirane novinske članke ($F1 = 32.8\%$), vidljiv je velik pad točnosti razrješavanja koreferencije.

Sustav koji su opisali Strube i Müller (2003) temelji se na nadziranom strojnom učenju primjenjenom na razrješavanje zamjenica u transkriptima dijaloga. Dijalog se može smatrati korakom između strukturiranih novinskih tekstova i šumovitog i nestrukturiranog korisnički generiranog sadržaja. Modele su trenirali na korpusu veličine tridesetak tisuća pojavnica, odnosno 3275 rečenica. Primjenom algoritma stabala odluke postigli su rezultate od $F1 = 47.42\%$ za sve tipove zamjenica, što pokazuje da je razrješavanje koreferencije puno teži zadatak već kad je u pitanju tekst dijaloga.

Jain et al. (1998) pozabavili su se razrješavanjem anafore u dijalozima više osoba. Oblikovali su sustav temeljen na znanju koristeći pristupe iz teorije grafove. Na nekoliko manjih korpusa sastavljenih od tekstova kazališnih drama i raznih dijaloga javljaju točnost između 62% i 83%.

Autori svih ovih sustava zaključuju da za uspješno rješavanje koreferencije u korisnički generiranom sadržaju nije dovoljno samo trenirati na dovoljno velikoj količini odgovarajućih tekstova, već je potrebno u oblikovanje sustava uklopiti i što veću količinu znanja o svijetu (engl. *world knowledge*), koje navode kao najveću prednost koju ljudi imaju prema računalima tijekom rješavanja ovakvih problema.

3. Problem razrješavanja koreferencije

Potpun se zadatak razrješavanja koreferencije entiteta sastoji od (1) početne obrade čistog teksta, zatim (2) otkrivanja potencijalnih spominjanja entiteta, a potom i (3) njihovog grupiranja po pojedinom entitetu. Ovaj je zadatak izuzetno semantičke prirode pa je i ljudima poprilično težak, što se najviše vidi u velikoj količini vremena potrebnoj za označavanje i u poprilično broju slučajeva u kojima se označivači ne slažu (detaljnije u (Ogrodniczuk et al., 2013)), a slično je primjećeno i prilikom označavanja korpusa za hrvatski jezik, korištenog u ovom radu. Koreferencija je fenomen koji se pojavljuje na razini dokumenta, odnosno svaki entitet na koji se neko spominjanje referencira definiran je na razini promatranog dokumenta.

Točan oblik zadatka rješavanog u ovom radu inspiriran je zadatkom s konferencije *CoNLL* 2011, opisanim u (Pradhan et al., 2011), ali i pristupima istraživača koji su radili ili rade na ovom problemu na jezicima srodnima hrvatskome, poput češkog, poljskog i bugarskog. Razlika u odnosu na zadatak rješavanja na CoNLL 2011 jest u tome što je u ovom radu zadatak ograničen samo na koreferencije među spominjanjima entiteta, gdje su spominjanja imeničke fraze. U zadataku s *CoNLL*-a promatrane su koreferencije među entitetima, ali i među događajima, odnosno i među entitetima i događajima. Također spominjanja nisu ograničena samo na imeničke fraze, već mogu biti i cijele surečenice ili rečenice (Pradhan et al., 2011).

3.1. Spominjanja

Imeničke fraze su fraze koje u rečenici imaju ulogu subjekta ili objekta, a kao glavni element tipično imaju imenicu ili zamjenicu. Tipovi imeničkih fraza koji su dio ovog zadatka uključuju:

- osobna imena i njihova proširenja (npr. „*Ivo Josipović*“, „*premijer Sanader*“, „*natjecatelj Ivo Ivić*“ i sl.);
- opće imenice s proširenjima („*predsjednik*“, „*manifestacija*“, „*mali, crni, krat-*

kodlaki pas“, „*taj korejski automobil*“ i sl.);

– i zamjenice („*on*“, „*njoj*“, „*ovo*“, „*taj*“ i sl.).

Dodatan problem predstavljaju ugniježđena spominjanja, poput „*predsjednik Republike Hrvatske*“, gdje je imenička fraza „*Republike Hrvatske*“ ugniježđena unutar navedene duže imeničke fraze. U ovom se radu ne obrađuju ugniježđena spominjanja, po uzoru na (Lee et al., 2011). Bitno je naglasiti da spominjanja ne prelaze granice rečenica, odnosno da se jedno spominjanje ne može protezati kroz više od jedne rečenice.

Kao što je ranije navedeno, ovaj je zadatak ograničen samo na spominjanja entiteta, dok se spominjanja događaja ne obrađuju. Primjerice u rečenici „*Njegov dolazak su roditelji dočekali s oduševljenjem.*“. Ovdje se imenička fraza „*Njegov dolazak*“ ne smatra spominjanjem jer se odnosi na događaj njegovog dolaska, dok bi se fraza „*roditelji*“ potencijalno mogla označiti kao spominjanje jer se referencira na entitet njegovih roditelja.

Opisanih imeničkih fraza u tekstu ima poprilično mnogo i njihova ekstrakcija nije naročito teška, pogotovo uz pomoć sintaktičkog analizatora, odnosno parsera. Glavni problem u ovom zadatku jest odrediti koje se od tih imeničkih fraza u tekstu referenciraju na neki određeni entitet iz stvarnog svijeta. Zatim je potrebno među tim spominjanjima pronaći ona koja tvore relaciju koreferencije s još barem jednim spominjanjem u promatranom dokumentu, odnosno pronaći entitete koje s spominje više od jednom u tekstu.

3.2. Relacija koreferencije

Relacija koreferencije je relacija ekvivalencije. Odnosno ona je:

– *Refleksivna* – svako spominjanje je koreferentno sa samim sobom;

– *Simetrična* – ako je prvo spominjanje u koreferenciji s drugim spominjanjem, onda je i drugo spominjanje u koreferenciji s prvim spominjanjem;

– *Tranzitivna* – ako je spominjanje A u koreferenciji sa spominjanjem B, i spominjanje B je u koreferenciji sa spominjanjem C, onda je i spominjanje A u koreferenciji sa spominjanjem C.

To implicira da relacija koreferencije vrijedni jednako među svim spominjanjima u dokumentu koja se odnose na isti entitet iz stvarnog svijeta i da redoslijed tih spominjanja nije bitan.

Dodatno, među koreferentnim spominjanjima može (ali i ne mora) vrijediti relacija anafore. Ova relacija vrijedi između dva spominjanja u tekstu i usmjerena je zdesna nalijevo, dok se istovjetna relacija koja vrijedi u obrnutom smjeru naziva katafora. Primjerice:

- Anafora – „***Sanjine*** *su fotografije izrazito cijenjene na ovim prostorima.* ***Ona*** *je jedan od najboljih hrvatskih fotografa.*“
- Katafora – „*Ljudi koji su* ***ga*** *cijenili, znali su da* ***Marko*** *nikad ne bi učinio takvo nešto.*“

U relaciji anafore interpretacija značenja krajnjeg spominjanja ovisi o prethodnom spominjanju, dok je u relaciji katafore obrnuto. Odnosno, ova se relacija može smatrati ograničenim oblikom relacije koreferencije. Međutim, postoje i parovi anaforičnih spominjanja koji nisu koreferentni.

Primjerice, u rečenici „*Po rukometašima koji sada nose dres Badel 1862 Zagreba* ***današnja momčad*** *ne bi smjela zaostajati za* ***onom*** *iz 1992. godine.*“ se značenje riječi „*onom*“ interpretira uz pomoć izraza „*današnja momčad*“, ali se one ne odnose na istu momčad, već na dvije različite momčadi odvojene u vremenu.

Također vrijedi i obrat, odnosno postoji i veliki broj koreferentnih parova spominjanja koji nisu u anaforičkom odnosu jer značenje jednog spominjanja ne ovisi o značenju drugog spominjanja, primjerice: „***Profesor Perić*** *se studentima obratio povodom Dana Fakulteta.* ***Dekan*** *je studentima poručio kako je ponosan na to što su postigli u protekloj godini.*“.

Detaljniji opis razlika između koreferencije i anafore dali su Van Deemter i Kibble (2000).

3.3. Vrste relacija koreferencije

Postoji nekoliko vrsta relacije koreferencije koje se mogu grupirati u pet kategorija:

- *Identitet* – odnosi se na sinonime, zamjenice i ostale izraze koji predstavljaju identičan identitet. Primjerice: „***Premijer Milanović*** *rekao je da* ***on*** *nikada nije odobrio taj zahtjev.*“;
- *Hiperonimija/hiponimija* – podrazumijeva parove spominjanja koja su u odnosu nadređenosti/podređenosti. Primjerice: „*Prije mjesec dana Ivan je kupio* ***novi automobil****.* ***Taj Mercedes*** *je čudo od auta.*“;
- *Metonimija* – predstavlja relaciju u kojoj se kao naziv entiteta koristi izraz čije je značenje drugačije, odnosno preneseno u danom kontekstu. Primjerice: „*Ju-*

čer su igrali ***Dinamo Zagreb*** *i Cibalija,* ***Zagrepčani*** *su slavili s ukupno tri pogotka.*“;

– *Meronimija* – podrazumijeva relaciju gdje jedno spominjanje predstavlja samo „dio od“ entiteta na koji se referencira drugo spominjanje. Primjerice: „*Od* ***jedanaestoro rukometaša****, danas su igrala samo* ***osmorica****.*“;

– *Nulta anafora* – odnosi se na specijalni slučaj koreferencije identiteta gdje je jedno od spominjanja skriveni subjekt. Skriveni subjekt je subjekt koji je neizrečen i može se pojaviti ili u surečenici ili u odvojenoj rečenici. U korpusu je kao to drugo anaforično spominjanje označen glagol koji predstavlja glavni element predikatnog skupa vezanog uz skriveni subjekt, primjerice: „***Marko*** *je išao u dućan.* ***Kupio*** *je deterdžent. Na povratku* ***sreo*** *je Ivana.*“.

Točnije, sve navedene vrste relacija osim identiteta smatraju se anaforičkim relacijama. Razlog tomu je što u svakoj od ovih relacija interpretacija značenja krajnjeg spominjanja ovisi o prethodnom spominjaju. Zbog toga posebice meronimija i metonimija ne predstavljaju klasičnu relaciju koreferencije jer ne predstavljaju relaciju ekvivalencije.

U ovom je radu obrađena samo relacija identiteta, po uzoru na zadatak s *CoNLL* 2011 (Pradhan et al., 2011), ali i s obzirom da se u dostupnom korpusu ne nalazi dovoljno velik broj primjera preostalih vrsta relacija. Bitno je naglasiti da to nije specifičnost korištenog korpusa, već se slični odnosi mogu vidjeti i u korpusima za druge jezike, poput (Ogrodniczuk et al., 2013).

4. Predobrada i označavanje skupa podataka

Skup podataka korišten u ovom radu sastoji se od novinskih članaka iz hrvatskih novina „Vjesnik“, iz razdoblja od 1999. do 2009. godine. U korpusu se nalaze članci raznolike tematike, od sporta, preko kulture i crne kronike, do unutarnjih i vanjskopolitičkih članaka. Novinski članci predstavljaju čiste, strukturirane tekstove, pisane standardnim jezikom. Zbog te činjenice, iz njih je moguće izvlačiti pouzdane i konzistentne jezične informacije, što nije slučaj u primjerice korisnički generiranom sadržaju.

4.1. Detalji označenog skupa

Statistički podaci o korištenom korpusu dani su u tablici 4.1. Ukupno je u korpusu označeno 15748 spominjanja i 12894 koreferentna para. Prosječan broj pojavnica po dokumentu je 555, prosječan broj spominjanja je 57, a prosječno je 47 koreferentnih parova. Iz tablice je vidljivo da spominjanja duljine do uključivo 4 pojavnice čine otprilike 92.5% svih spominjanja u korpusu, dok se najdulje označeno spominjanje sastoji od ukupno 27 pojavnica.

Također je vidljivo da je koreferencija identiteta daleko najbrojnija vrsta relacije koreferencije i čini otprilike 87% svih koreferentnih parova u korpusu, što opravdava odluku ograničavanja na samo tu vrstu koreferencije u ovom radu. Zanimljivo je primjetiti da se u prosjeku po dokumentu govori o 13 različitih entiteta iz stvarnog svijeta, od kojih se svaki od njih u prosjeku spominje 4 puta u danom dokumentu.

4.2. Označavanje skupa podataka

Izrada ovog korpusa motivirana je činjenicom da za hrvatski jezik nije postojao zadovoljavajuće velik i kvalitetan korpus s označenim koreferencijama. Označavanje je

Tablica 4.1: Statistike korpusa

Obilježje	Vrijednost
Članaka/dokumenata	265
Pojavnica	146998
Jedinstvenih riječi	27853
Jedinstvenih lema	17518
Spominjanja duljine 1 pojavnice	8657
Spominjanja duljine 2 pojavnice	3814
Spominjanja duljine 3 pojavnice	1366
Spominjanja duljine 4 pojavnice	735
Spominjanja duljine 5-27 pojavnica	1176
Koreferencija identiteta	11239
Koreferencija nadopojma/podpojma	64
Koreferencija metonimije	112
Koreferencija meronimije	922
Koreferencija nulte anafore	557
Entiteta	3699

provedeno neposredno prije početka oblikovanja ovog sustava. Cjelokupni projekt označavanja trajao je oko tri mjeseca i na njemu je radilo šest označivača. Razvijene su detaljne upute za označavanje sukladno ranije opisanom pristupu rješavanja problema koreferencije u tekstovima na hrvatskome jeziku (v. poglavlje 3).

4.2.1. Obuka i kalibracija

Označivači su obučeni na sastancima kroz diskusiju o primjerima, ali i pomoću dva kalibracijska skupa od po 17 dokumenata. Označavanje je provedeno koristeći prilagođeni alat za označavanje s grafičkim sučeljem, originalno razvijen u okviru označavanja korpusa imenovanih entiteta za sustav od Glavaš et al. (2012). Za označavanje kalibracijskih skupova označivačima je u prosjeku trebalo oko pet sati po skupu.

4.2.2. Označavanje dokumenata

Označavanje je provedeno u dva dijela, gdje su označivači označavali skupove dokumenata u parovima. Označavanje u parovima odabrano je kako bi se podigla kvaliteta označenog teksta, zbog toga što različiti označivači imaju različite perspektive i stilove označavanja. Bez obzira na iznimno detaljne i precizne upute, što je posljedica težine ovog zadatka. Parovi označivača za prvi skup dokumenata bili su različiti od parova koji su označavali drugi skup dokumenata. Svaki je par označivača dobio otprilike 45 dokumenata na označavanje, što uz tri para označivača po skupu daje ukupno 265 označenih dokumenata.

Za označiti jedan skup od 45 dokumenata, označivačima je u prosjeku trebalo 12h, s time da im je dano otprilike tjedan dana vremena za svaki skup. Označivači su po završetku označavanja naveli da ovaj zadatak zahtijeva vrlo visoku razinu koncentracije i velik kognitivni napor te da je ovako nešto moguće precizno raditi u neprekidnom trajanju od najviše tri sata.

Ovakva iskustva imaju smisla bez obzira na činjenicu da ljudi razrješavanje koreferencija rade nesvjesno i automatski, i to svakodnevno, odnosno kad god čitaju neki tekst ili slušaju nekoga kako govori. U takvim uobičajenim okolnostima nije potrebno toliko precizno i detaljno, a ni u potpunosti svjesno, odrediti što je točno u koreferenciji s čim, kao što je to slučaj u ovom zadatku, već takve spoznaje dolaze „same od sebe“ kao dio ljudskog procesa razumijevanja prirodnog jezika.

4.2.3. Razrješavanje neslaganja

Kako bi se označeni korpus dodatno pročistio, po završetku označavanja autor ovog rada prošao je kroz cijeli skup dokumenata i proveo razrješavanje neslaganja označivača. Na razrješavanje je utrošeno ukupno oko 25 sati rada, u rasponu od nekoliko tjedana. Razrješavanje je također provedeno uz pomoć prilagođenog alata s grafičkim sučeljem, također originalno razvijenog za ranije spomenuti korpus imenovanih entiteta.

Označivači se u parovima u prosjeku ne slažu u 70% svih označenih koreferentnih parova u danom skupu dokumenata. Od toga je u prosjeku u 24% slučajeva problem u drugačijem uparivanju spominjanja, odnosno jedno spominjanje u paru dijele, ali drugo je različito. Zatim je u otprilike 5% slučajeva problem u neslaganju oko točnih granica spominjanja, ali se spominjanja u koreferentnom paru bar djelomično poklapaju. I za kraj se označivači u otprilike 1% neslaganja ne slažu samo u tipu koreferencijskog odnosa, dok su označena spominjanja jednaka.

Ostala se neslaganja odnose ili na lažno koreferentna spominjanja, odnosno da je jedan označivač pogrešno zaključio da se radi o koreferenciji, ili se jedan označivač naprosto propustio označiti neki koreferentni par. Ovako velik broj neslaganja opravdava potrebu za ručnim razrješavanjem.

Na taj je način korpus sveden na najvišu moguću razinu točnosti, s obzirom na raspoložive resurse, i može se smatrati zlatnim standardom za hrvatski jezik. Valja napomenuti da je potrebno uzeti u obzir težinu ovog zadatka za ljude i da čak i nakon razrješavanja u korpusu postoje rijetki slučajevi za čije točno označavanje nije bilo moguće postići konsenzus. Tu se uglavnom radi o rjeđim vrstama relacije koreferencije, a ne o identitetu.

5. Sustav za razrješavanje koreferencije

Sustav opisan u nastavku razvijen je kao cjelovito rješenje (engl. *end-to-end*) zadatka razrješavanja koreferencije u tekstovima na hrvatskome jeziku. Sustav na ulaz prima čiste tekstne dokumente i kao izlaz vraća te iste dokumente s označenim koreferentnim spominjanjima, u formatu SemEval-2010.[1]

Struktura sustava inspirirana je cjevovodom obrade prirodnog jezika predloženim od Soon et al. (2001), s određenim razlikama s obzirom na dostupnost jezičnih alata za hrvatski jezik. Sustav oblikuje vektor značajki za svako spominjanje, odnosno za svaki par spominjanja. Potom se provodi klasifikacija po modelu parova spominjanja (engl. *mention-pair model*) koristeći metode statističkog nadziranog strojnog učenja.

5.1. Oblikovanje vektora značajki

Sustav uzima čiste tekstne dokumente koje rastavlja na rečenice i tokenizira te pritom provodi osnovnu morfološku obradu i čisti tekst od potencijalnih nepotrebnih specijalnih znakova. Tako obrađeni dokumenti se predaju označivaču vrsta riječi i lematizatoru od Agić et al. (2013). Dokumenti se potom provlače kroz ovisnosni sintaktički analizator, odnosno parser, za hrvatski jezik od Agić i Merkler (2013).

Bez obzira na dostupnost visokokvalitetnog alata za ekstrakciju imenovanih entiteta iz tekstova na hrvatskome jeziku, *CroNER*, opisanog u (Glavaš et al., 2012), odlučili smo se u ovoj fazi ne iskoristiti taj alat pri izradi ovog sustava. Ta je odluka donesena iz razloga što je *CroNER* treniran upravo na istom korpusu novinskih tekstova kao i ovaj sustav. Zbog te činjenice bi korištenje alata *CroNER* donijelo nepravednu prednost pri ekstrakciji imenovanih entiteta, koji su po granicama u pravilu slični ili čak istovjetni potencijalnim spominjanjima.

[1] `http://stel.ub.edu/semeval2010-coref/datasets#formatting`

Nakon toga se iz teksta izvlače imeničke fraze koristeći ručno izrađenu metodu temeljenu na pravilima. Takvim se imeničkim frazama heuristički određuje rod i broj, koji se potom koriste u odabiru parova kandidata spominjanja. Odvojeno se provodi i ekstrakcija označenih podataka za treniranje. Nakon toga se izračunava Levenshteinova udaljenost između dva spominjanja i kodira se završni oblik vektora značajki. U nastavku je dan detaljniji pregled navedenih koraka.

5.1.1. Tokenizacija i rastavljanje na rečenice

Kao prvi korak u obradi tekstnih dokumenata, provodi se tokenizacija i rastavljanje na rečenice. Pod tokenizacijom se podrazumijeva rastavljanje teksta na riječi i interpunkciju, s time da se interpunkcijski znakovi također čuvaju kao zasebne pojavnice. Za svaku se pojavnicu također pamti njezina početna pozicija, odnosno broj znakova od početka dokumenta, što se kasnije koristi u ekstrakciji podataka za treniranje. Tokenizacija je izvedena heuristički, koristeći jednostavne regularne izraze, uz ranije spomenuti uvjet očuvanja interpunkcijskih znakova.

Rastavljanje teksta na rečenice izvedeno je ručno rađenom metodom temeljenom na pravilima. Ova metoda, osim što rastavlja tekst na riječi, također rekonstruira datume, vremena i redne brojeve u tekstu. Metoda koristi ranije pripremljenu listu poznatih kratica u kombinaciji s pretpostavkom o pouzdanoj informaciji o velikom početnom slovu riječi koje slijede nakon točke. Metoda slijedno prolazi kroz tekst i uvijek promatra najviše jednu pojavnicu unatrag pri donošenju odluke o granici rečenica.

5.1.2. Označavanje vrsta riječi i lematizacija

Nad tako tokeniziranim dokumentima provedeno je označavanje vrsta riječi pomoću alata otvorenog koda *Hunpos*[2] i lematizacija[3] pomoću alata također otvorenog koda *CST's Lemmatiser*[4]. Za oba alata razvijeni su modeli za hrvatski jezik, opisani u (Agić et al., 2013). Oba se alata temelje na metodama nadziranog statističkog strojnog učenja, s time da je alat za označavanje vrsta riječi temeljen na klasifikatoru skrivenih Markovljevih modela (engl. *hidden markov models, HMM*), a lematizator na automatskom učenju pravila.

Označivač vrsta riječi podržava punu specifikaciju MULTEXT East[5] verzije 5 za

[2] `https://code.google.com/p/hunpos/`

[3] Lematizacija podrazumijeva svođenje riječi na osnovni, rječnički oblik.

[4] `http://cst.dk/online/lemmatiser/uk/`

[5] `http://nl.ijs.si/ME/Vault/V3/msd/html/msd.html`

Tablica 5.1: Morfosintaktičke kategorije

Vrsta riječi	Kod	Atributa
Imenica	N	10
Glagol	V	15
Pridjev	A	12
Zamjenica	P	17
Prilog	R	6
Prijedlog	S	4
Veznik	C	7
Broj	M	12
Usklik	I	2
Ostalo	X	0
Skraćenica	Y	5
Čestica	Q	3

hrvatski jezik. Vrste riječi podržane ovom specifikacijom navedene su u tablici 5.1, zajedno s brojem atributa svake morfosintaktičke oznake (engl. *morphosyntactic description, MSD*). Iz MSD oznaka za sve se imenice, pridjeve i zamjenice ekstrahira posebno rod i broj, koji se kasnije koriste u heurističkom određivanju roda i broja imeničke fraze.

5.1.3. Ovisnosna sintaktička analiza

Nad lematiziranim korpusom, s označenim vrstama riječi, provodi se ovisnosna sintaktička analiza pomoću alata otvorenog koda *MSTParser*[6]. Za navedeni alat razvijen je model za hrvatski jezik, opisan u (Agić i Merkler, 2013). Alat se temelji na algoritmu najmanjeg razapinjajućeg stabla (engl. *minimum spanning tree, MST*).

Ovisnosni parser (engl. *dependency parser*) gradi sintaktičko stablo u kojem su označene sintaktičke ovisnosti među riječima prema SETimes banci ovisnosnih stabala za hrvatski jezik. Podržane vrste ovisnosti među riječima navedene su u tablici 5.2. Svaka riječ ima točno jednog roditelja o kojem ovisi, dok se na vrhu sintaktičkog stabla nalazi riječ koja predstavlja glavu predikata i o kojoj ovise atributi predikata, ali i subjekt.

[6] `http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html`

Tablica 5.2: Vrste sintaktičkih ovisnosti, prema Agić et al. (2013)

Vrsta	Kod
Prilog	Adv
Apozicija	Ap
Atribut	Atr
Predikatni proširak	Atv
Pomoćni glagol	Aux
Koordinacijski veznik	Co
Elipsa	Elp
Objekt	Obj
Ostalo	Oth
Imenski predikat	Pnom
Predikat	Pred
Prijedlog	Prep
Interpunkcija	Punc
Subjekt	Sb
Subordinacijski veznik	Sub

5.1.4. Ekstrakcija imeničkih fraza

Ekstrakcija imeničkih fraza nastupa nakon što se prikupe podaci za svaku pojavnicu iz dostupnih jezičnih alata. Ova je metoda zasnovana na ručno oblikovanim pravilima koja u najvećoj mjeri u obzir uzimaju samo vrstu riječi. Razlog tomu je da su prethodno opisani napredniji jezični alati, poput ovisnosnog parsera, postali dostupni tek nedavno pa nije bilo vremena za uklopiti nove informacije proizašle iz tih alata u oblikovanje ove metode.

Metoda slijedno prolazi kroz tekst i traži sve imenice, zamjenice, pridjeve i brojevne izraze kao početne pojavnice imeničke fraze. Metoda dodaje pojavnice koje su u skupu ovih vrsta riječi dok prvi put ne naiđe na pojavnicu koja je van tog skupa. U tom trenutku još jednom provjerava zadnji zabilježenu pojavnicu u netom završenoj imeničkoj frazi i izbacuje određene vrste zamjenica (poput *taj, neki, ova, koje* i sl.), ali i brojevne izraze ako se nalaze na tom zadnjem mjestu. Nakon toga metoda kreće ispočetka i traži prvog sljedećeg kandidata koji započinje novu imeničku frazu. Primjer ekstrahiranih imeničkih fraza iz jedne rečenice dan je u nastavku.

„*Za [našu javnost], u kojoj se često [erudicija] i [načitanost] brkaju s [brbljavošću] i [galamom], a kojoj je [glavna tema ulazak] u [Europsku uniju], posebno bi mogla biti intrigantna [Papina razmišljanja] o [Europi], njeni [sadašnji] i [budući temelji].*“

Iz primjera je vidljivo da u metodi itekako ima prostora za poboljšanje, primjerice korištenjem podataka o sintaktičkim ovisnostima među riječima, ali u principu metoda postiže svoj osnovni cilj, a to je visok odziv pri ekstrakciji imeničkih fraza. Također jedno od bitnih budućih poboljšanja jest i podrška za ugniježđene imeničke fraze, poput „*[predsjednik [Republike Hrvatske]]*“. Ovaj se problem rješava upotrebom sintaktičkog parsera ili ako isti nije dostupan, razdjelnikom (engl. *chunker*) koji podržava ugniježđene fraze.

5.1.5. Ekstrakcija označenih podataka za treniranje

Ovaj problem, premda na prvi pogled trivijalan, ipak nije tako jednostavan. Spominjanja su u tekstu označena slobodno, jer tekst nije prethodno tokeniziran, i za svako od tih spominjanja potrebno je formirati vektore značajki, što sustav izvodi automatski. Soon et al. (2001) predložili su pristup da, ako sustav ne ekstrahira imeničku frazu koja se u granicama ne podudara u potpunosti s označenim spominjanjem, primjer za učenje se ne formira. Odnosno, ako je sustav prepoznao imeničku frazu koja predstav-

lja samo dio označenog spominjanja ili takvu imeničku frazu uopće nije označio, onda se to spominjanje ne koristi u oblikovanju parova za treniranje.

Ovakav pristup postavlja velika očekivanja na odziv metode za ekstrakciju imeničkih fraza iz teksta, što predstavlja rizik neiskorištavanja zamjetnog broja primjera za treniranje. U ovom je sustavu iskorišten modificirani pristup koji koristi označena spominjanja za ekstrakciju imeničkih fraza i tako ekstrahirane pozitivne primjere dodaje skupu automatski ekstrahiranih imeničkih fraza.

Na ovaj je način moguće iskoristiti veliku većinu (pozitivnih) primjera za učenje, a automatski ekstrahirane imeničke fraze u tom slučaju pak služe za generiranje negativnih primjera za učenje. I dalje nije nužno slučaj da će se ekstrahirati baš svi označeni parovi, što proizlazi iz nesavršenosti tokenizacije, ali broj takvih neiskorištenih spominjanja je na ovaj način sveden na najmanju moguću mjeru (ukupno 170, što je manje od jednog spominjanja po dokumentu). Tu se uglavnom radi o problemu znakova koji nisu dio abecede hrvatskoga jezika, a kakvi se najčešće pojavljuju u stranim imenima, poput „Schröeder“ i sl. Tokenizator takva slova tretira kao granicu između riječi te takva imena greškom razdvaja na dvije riječi.

Svi dohvaćeni koreferentni parovi grupiraju se po entitetima. Ako je svako spominjanje čvor, a svaka koreferencija brid koji spaja dva čvora, onda se dobiveni graf jednog entiteta može predočiti kao stablo. Svi entiteti unutar jednog dokumenta onda čine šumu. Ako se prihvati konvencija da je za svako spominjanje unutar jednog entiteta njegov antecedent ono spominjanje (istog entiteta) koje mu je najbliže u tekstu slijeva, onda se entitet može predočiti kao lanac. Pozitivni parovi za učenje se odabiru tako da se generira svaki mogući par spominjanja unutar jednog entiteta.

5.1.6. Odabir parova kandidata spominjanja

U idealnom bi slučaju, prema modelu parova spominjanja, trebalo provesti klasifikaciju svakog para spominjanja u dokumentu. To bi značilo klasifikaciju reda veličine n^2 parova po dokumentu, gdje je *n* broj ekstrahiranih spominjanja u dokumentu. U ovom je sustavu korišten pristup koji koristi kombinaciju triju heuristika kako bi smanjio broj parova za klasifikaciju. Motivacija za ovaj pristup dana je ranije u poglavlju koje opisuje model parova spominjanja.

Prva određuje neki najveći, relativno malen broj parova koje neki spominjanje može tvoriti u tekstu. Parovi se formiraju tako da se uzme jedno spominjanje i uparuje ga se sa svim spominjanjima koja su bila prije njega u tekstu, sve dok se ne dosegne navedena granica broja parova.

Druga heuristika definira da se uparena spominjanja moraju slagati u rodu i broju, što predstavlja dodatni filter na prethodno dobivene parove. Ovakva je heuristika opravdana činjenicom da svaki entitet može imati samo jedan rod i broj pa je dozvoljeno pretpostaviti da će njegova spominjanja imati upravo takav rod i broj. Primjerice, pretpostavlja se da će svako spominjanje predsjednika Josipovića biti muškog roda jednine.

Treća heuristika govori da dvije zamjenice ne mogu tvoriti koreferentni par. To je motivirano činjenicom da dvije zamjenice same za sebe ne nose dovoljno informacija o interpretaciji njihovog značenja.

Rod i broj nekog spominjanja se također određuje heuristički, pomoću roda i broja pojavnice. Pretpostavlja se da prva pojavnica, kojem je označivač vrsta riječi uspio dodijeliti rod i broj, predstavlja rod i broj cijelog spominjanja. Primjerice, u spominjanju „*veliki bijeli pas krivih nogu*“ pretpostavlja se da muški rod jednine prve riječi „*veliki*“ predstavlja rod i broj kompletnog spominjanja, dok je manje vjerojatno da ženski rod množine riječi „*nogu*“ daje točnu procjenu roda i broja cijelog spominjanja.

5.1.7. Računanje Levenshteinove udaljenosti

Ideja korištenja Levenshteinove udaljenosti između dvaju spominjanja u potencijalno koreferentnom paru preuzeta je od Strube et al. (2002). Levenshteinova udaljenost je mjera udaljenosti između dva znakovna niza, predložena od Levenshtein (1966). Definirana je kao najmanji broj potrebnih umetanja, brisanja ili zamjene jednog znaka kako bi se jedan znakovni niz transformirao u drugi. Zamjena jednog znaka definira se kao jedno brisanje i jedno umetanje pa se broji kao dvije promjene.

Ovdje primjenjeni algoritam računanja Levenshteinove udaljenosti opisan je u (Bellman i Dreyfus, 1962) i primjer je algoritma dinamičkog programiranja. Primjerice, Levenshteinova udaljenost između riječi „*predsjednik*“ i „*predsjednički*“ jest 2, a između izraza „*moj pas*“ i „*moja mačka*“ je 7. Udaljenost između dvaju izraza govori koliko su oni slični i trebala bi poslužiti kao korisna heuristika u odluci o potencijalnoj koreferenciji između dvaju spominjanja.

5.1.8. Oblikovanje završnih vektora značajki

U ovom se koraku provodi završno kodiranje vrijednosti značajki vektora za klasifikaciju. Korišteni klasifikator zahtijeva uporabu isključivo binarnih značajki pa je potrebno sve multinomijalne značajke pretvoriti u binarne. Ključno je osigurati konzistentnost u kodiranju vrijednosti značajki neovisno radi li se o učenju ili predikciji.

Iz tog se razloga tijekom pripreme skupa za učenje spremaju jedinstvene vrijednosti svih multinomijalnih značajki, uključujući i riječi i leme. Ako se kasnije tijekom predikcije pojavi riječ ili lema koja se nije prethodno pojavila u skupu za učenje, ista biva kodirana rezerviranom vrijednošću značajke za nepoznatu vrijednost. Odnosno, klasifikator se ne može uspješno nositi s vrijednostima značajki koje nije prethodno susreo, već ih kodira pomoću pretpostavljenih vrijednosti.

Završni oblik vektora značajki sadrži sljedeće značajke, za svako od dvaju spominjanja zasebno:

- *Lema*;
- *Oznaka vrste riječi zajedno s MSD značajkama*;
- *Vrsta sintaktičke ovisnosti*;
- *Rod spominjanja*;
- *Broj spominjanja*;
- *Levenshteinova udaljenost između dvaju spominjanja*.

Prve tri značajke izračunavaju se za svaku pojavnicu svakog od dvaju spominjanja, što čini ukupno 8 vrijednosti svake od tih triju značajki po vektoru. Spominjanja su ograničena na do četiri pojavnice, što (kako je ranije spomenuto) u potpunosti pokriva oko 92.5% svih spominjanja u skupu za učenje, dok se za dulja spominjanja preostale pojavnice ignoriraju.

Značajke za rod i broj spominjanja dane su samo jednom jer se parovi spominjanja filtriraju upravo po slaganju u rodu i broju. Levenshteinova se udaljenost računa promatrajući oba kompletna spominjanja, a ne samo prve četiri pojavnice.

5.2. Klasifikacija i grupiranje spominjanja

Vektori značajki oblikovani na gore opisani način klasificiraju se pomoću treniranog modela stroja s potpornim vektorima. Zatim se izlaz klasifikatora dekodira, provodi se grupiranje spominjanja po entitetima i dokumenti se prebacuju u ranije spomenuti format SemEval-2010. U nastavku je dan detaljniji pregled navedenih koraka.

5.2.1. Klasifikacija parova spominjanja

Klasifikacija parova spominjanja izvedena je pomoću prethodno naučenog modela stroja s potpornim vektorima. Klasifikator na ulaz prima pripremljeni skup kodira-

nih vektora značajki i model, a na izlaz vraća predviđene oznake za svaki dani vektor značajki.

5.2.2. Grupiranje spominjanja po entitetima

Grupiranje spominjanja po entitetima izvedeno je analogno postupku u ekstrakciji podataka za učenje. Svaki koreferentni par spominjanja čini jedan brid stabla koje predstavlja entitet. Stablo se generira uzastopnim spajanjem koreferentnih parova i skupova koji sadrže zajednička spominjanja. Na ovaj se način dobivaju razredi ekvivalencije, koristeći pretpostavku o tranzitivnosti relacije koreferencije.

5.2.3. Konverzija u format SemEval-2010

Završno se cjelokupni skup podataka zapisuje u formatu SemEval-2010[7], koji je čitljiv čovjeku, ali je ujedno i zahtijevani format korištenog evaulatora. Spomenuti format sadrži svaku pojavnicu u zasebnom retku, zajedno s pripadnim vrijednostima značajki. Granice spominjanja su definirane okruglim zagradama na početnoj i završnoj pojavnici uz dodatak jedinstvenog cjelobrojnog identifikatora entiteta kojem pripada to spominjanje, u promatranom dokumentu.

U nastavku je dan primjer jedne rečenice iz skupa za učenje, zapisane u navedenom formatu.

```
1 Sanader sanader N-msn rekao atribut (1)
2 je biti Vcr3s rekao atribut _
3 rekao reći Vmp-sm [root] predikat _
4 kako kako Cs rekao subordinacijski_veznik _
5 je biti Vcr3s kako predikat _
6 činjenica činjenica N-fsn je imenski_predikat _
7 da da Cs činjenica subordinacijski_veznik _
8 je biti Vcr3s da predikat _
9 Unija unija N-fsn je imenski_predikat (6)
10 umorna umorna N-fsn Unija atribut _
11 od od Sg umorna prijedlog _
12 proširenja proširenje N-nsg od atribut _
13 ,,Z da interpunkcija _
14 da da Cs kako subordinacijski_veznik _
```

[7] http://stel.ub.edu/semeval2010-coref/datasets#formatting

```
15 ima imati Vmr3s da predikat _
16 problema problem N-mpg ima objekt _
17 s s Si problema prijedlog _
18 Lisabonskim lisabonski Agpmsi ugovorom atribut (5
19 ugovorom ugovor N-msi s atribut 5)
20 ,,Z a interpunkcija _
21 a a Cs i atribut _
22 i i Cc ima koordinacijski_veznik _
23 financijska financijski Agpfsn kriza atribut (13
24 kriza kriza N-fsn pogodila atribut 13)
25 je biti Vcr3s pogodila atribut _
26 pogodila pogoditi Vmp-sf i atribut _
27 Europu europa N-fsa pogodila atribut (12)
28 . . Z [root] interpunkcija _
```

6. Programsko ostvarenje

Opisani sustav implementiran je u programskom jeziku *Java* verzije 7. Sustav je oblikovan kao desktop aplikacija predviđena za pokretanje na platformi *Microsoft Windows*. Klasifikator korišten za izradu modela i predikciju jest LibSVM[1] opisan u (Chang i Lin, 2011). Za vrednovanje sustava iskorišten je službeni evaluator[2] verzije 4, konferencije *CoNLL* 2011 (engl. *Conference on Computational Natural Language Learning 2011*), opisan u (Recasens et al., 2010).

S obzirom da se cijeli sustav temelji na nadziranom statističkom strojnom učenju, u nastavku je dan detaljniji opis modela stroja s potpornim vektorima, koji je korišten u ovom sustavu, s naglaskom na verziju implementiranu u alatu LibSVM.

6.1. Stroj s potpornim vektorima

Stroj s potpornim vektorima (engl. *support vector machine, SVM*) vrlo je popularan diskriminativni model za klasifikaciju s primjenama u mnogim područjima (Hsu et al., 2010). Model je u današnjem obliku predložen od Cortes i Vapnik (1995) i temelji se na rješavanju optimizacijskog problema kvadratnog programiranja.

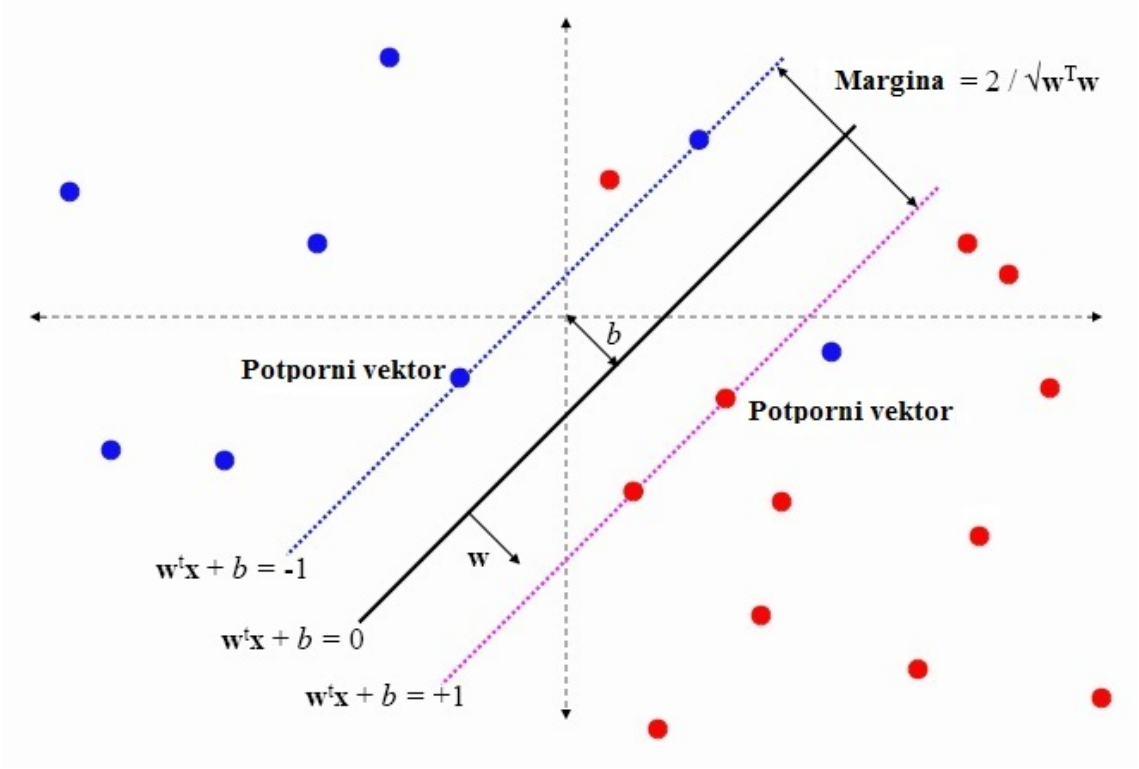
Ako je dan skup za učenje s parovima instanci i oznaka $(x_i, y_i), i = 1, ..., l)$, gdje je $x_i \in \mathbb{R}^n$ i $y \in \{1, -1\}^l$, onda je potrebno naći rješenje problema:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ &y_i(\mathbf{w}^T x_i + b) \geqslant 1 - \xi_i, \\ &\xi_i \geqslant 0. \end{aligned} \tag{6.1}$$

SVM pronalazi hiperravninu s najvećom marginom koja linearno razdvaja pozitivne od negativnih primjera u tom prostoru. Hiperravnina se zapisuje koristeći pot-

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[2] `http://conll.cemantix.org/2011/software.html`



Slika 6.1: Vizualizacija margine SVM-a

porne vektore, odnosno primjere za učenje koji se nalaze na pronađenoj margini. Vizualizacija ovog koncepta vidljiva je na slici 6.1[3]. Parametar $C > 0$ označava kaznu za greške u klasifikaciji, odnosno kažnjava primjere za učenje koji se nalaze sa suprotne strane margine.

Boser et al. (1992) predložeili su tzv. jezgrene funkcije koje mapiraju vektore za učenje u prostor s više dimenzija, koristeći (potencijalno) nelinearne transformacije. Popularne jezgrene funkcije, osim linearne, su i polinomijalna, radijalna i sigmoidalna. U visokodimenzionalnim prostorima prihvatljivo je prepostaviti linearnu razdvojivost klasifikacijskog problema pa nije nužno koristiti jezgrene funkcije.

U slučaju linearnog SVM-a poželjno je optimirati po vrijednosti parametra *C*. Odnosno, poželjno je korištenjem n-terostruke unakrsne provjere koju nudi LibSVM provesti optimizaciju vrijednosti parametra *C*. Kao što je ranije navedeno, ovaj parametar predstavlja kaznu za pogrešno klasificirane primjere. Vrijednosti koje se tipično isprobavaju su 2^{-5} do 2^{15} u koracima od jedne potencije (Hsu et al., 2010) te se na kraju model trenira s vrijednošću koja je tijekom optimizacije proizvela najveću točnost klasifikacije.

[3]Slika lokalizirana i preuzeta s `http://www.ifp.illinois.edu/~yuhuang/sceneclassification.html`.

ako se koristi neka od ranije spomenutih jezgrenih funkcija, onda se optimizacija vrijednosti parametara provodi i za parametar te funkcije. Izvodi se tzv. rešetkasta pretraga (engl. *grid search*) gdje se isprovaju različite kombinacije vrijednosti oba parametra te se na kraju model trenira s najboljim pronađenim vrijednostima.

7. Vrednovanje uspješnosti

7.1. Evaluacijske mjere

Za vrednovanje uspješnosti postupka iskorišten je službeni evaluator[1] konferencije *CoNLL* 2011 (engl. *Conference on Computational Natural Language Learning 2011*). Navedeni evaluator koristi četiri mjere uspješnosti: *MUC*, *CEAF*, *B-CUBED*, *BLANC*, koje su opisane u nastavku. Klasični (netežinski) prosjek triju tih mjera, *MUC*, *B-CUBED* i *CEAF*, uzima se kao ukupna službena mjera za evaluaciju.

Za skup entiteta $\mathcal{K}$ iz ručno označenog skupa i za skup entiteta $\mathcal{R}$ iz skupa koji je označio sustav koji je potrebno evaluirati, svaka od mjera generira svoju varijaciju mjera preciznosti i odziva, odnosno F1.

7.1.1. Preciznost, odziv i F1

S obzirom da preostale mjere uspješnosti koriste svoje varijante mjera preciznosti i odziva, potrebno je prvo dati definiciju ovih mjera u općenitom slučaju.

Preciznost je mjera koja se definira kao omjer između točnih pozitivnih primjera (engl. *true positive, TP*) i zbroja TP i lažno pozitivnih primjera (engl. *false positive, FP*). Odziv se pak definira kao omjer broja TP i zbroja TP i lažno negativnih primjera (engl. *false negative, FN*). F1 mjera je harmonijski prosjek ove dvije mjere. Odnosno:

$$Preciznost = \frac{TP}{TP + FP}$$

$$Odziv = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Preciznost * Odziv}{Preciznost + Odziv}$$

[1] `http://conll.cemantix.org/2011/software.html`

7.1.2. *MUC*

Mjera *MUC* (engl. *Message Understanding Conference*) predložena je u (Vilain et al., 1995) i najstarija je i najčešće korištena mjera uspješnosti u problemu razrješavanja koreferencije entiteta.

Mjera je usredotočena na veze, odnosno parove spominjanja. Odziv predstavlja omjer broja veza između entiteta u skupovima $\mathcal{K}$ i $\mathcal{R}$, i broja veza u skupu $\mathcal{K}$. Preciznost je pak definirana kao omjer broja veza između entiteta u skupovima $\mathcal{K}$ i $\mathcal{R}$, i broja veza u skupu $\mathcal{R}$. Ova mjera preferira sustave koji imaju više spominjanja po entitetu, a sustav koji sva spominjanja spoji u jedan entitet dobiva najveći odziv bez nekog značajnog pada vrijednosti preciznosti.

7.1.3. CEAF

CEAF (engl. *Constrained Entity Allignment F-Measure*) je mjera predložena u (Luo, 2005) i za nju postoje dvije varijacije, jedna fokusirana na entitete i druga fokusirana na spominjanja. Ovaj evaluator koristi mjeru usredotočenu na entitete.

Ova mjera poravnava svaki entitet iz skupa $\mathcal{R}$ s najviše jednim entitetom iz skupa $\mathcal{K}$ tako da traži najbolje jedan-na-jedan preslikavanje između entiteta koristeći matricu sličnosti. Odziv se računa tako da se ukupna sličnost podijeli s brojem spominjanja u skupu $\mathcal{K}$, a preciznost je pak omjer ukupne sličnosti s brojem spominjanja u skupu $\mathcal{R}$.

7.1.4. B-CUBED

B-CUBED mjera je predložena od Bagga i Baldwin (1998) i pokušava riješiti neke probleme koji postoje u *MUC* mjeri.

Ova se mjera koncentrira na spominjanja i izračunava odziv i preciznost za svako spominjanje. Ako je K entitet iz ručno označenog skupa i sadrži spominjanje M, a R je entitet iz automatski označenog skupa i sadrži spominjanje M, onda se odziv za spominjanje M izračunava kao omjer veličine presjeka K i R i veličine entiteta K, u smislu broja spominjanja po entitetu. Preciznost se pak računa kao omjer veličine presjeka K i R i veličine entiteta R. Ukupni odziv i preciznost definirani su kao prosjek pojedinih odziva i preciznosti za svako spominjanje.

7.1.5. BLANC

Mjera *BLANC* (engl. *BiLateral Assessment of Noun-phrase Coreference*) je predložena od Recasens i Hovy (2011) i koristi varijaciju tzv. *Randovog indeksa* predloženog od

Rand (1971). *Randov indeks* je u ovom slučaju iskorišten kao mjera sličnosti između dvije grupe podataka.

BLANC za računanje odziva koristi klasičnu varijantu odziva posebno za koreferentne parove i posebno za nekoreferentne parove, a zatim kao ukupni odziv uzima njihov prosjek. Preciznost se računa analogno, s time da se ovaj put promatraju klasične varijante formula preciznosti. Ukupna preciznost i ukupna vrijednost F1 mjere se računaju kao prosjek odgovarajućih iznosa za svaki par spominjanja.

7.2. Eksperimenti

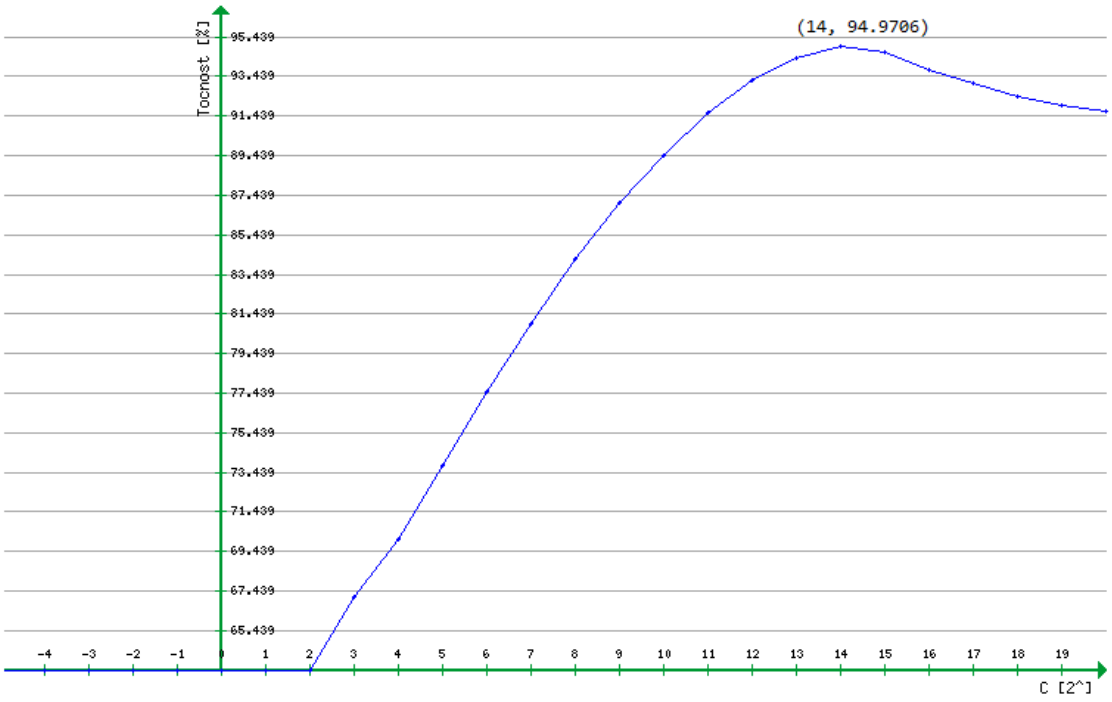
U ovom je poglavlju dan pregled provedenih eksperimenata u svrhu vrednovanja uspješnosti implementiranog postupka. Dani su rezultati eksperimenata zajedno s analizom pogrešaka.

7.2.1. Vrednovanje oblikovanja primjera za učenje

Kreiranje primjera za učenje započinje ekstrakcijom potencijalnih spominjanja iz teksta. Implementirana metoda temeljena na ručno oblikovanim pravilima na ručno označenim spominjanjima iz ranije opisanog korpusa (v. poglavlje 4) postiže odziv od 50.3% i preciznost od 22.7%. Ovakvi rezultati jasan su pokazatelj potrebe poboljšanja navedene metode. Točnije, niska preciznost u ovom slučaju nije toliko bitna jer su u korpusu označavana samo spominjanja koja tvore barem jedan koreferentni par.

Metoda trenutno uzima u obzir samo oznake vrsta riječi pojavnica u kombinaciji s listom neželjenih tipova pojavnica, kako je objašnjeno ranije. Ograničenja u oblikovanju ove metode glavni su razlog niskog odziva. Tu se najčešće radi o tome da metoda prihvaća samo imenice, pridjeve i zamjenice kao potencijalne dijelove spominjanja, što u općenitom slučaju nije valjana pretpostavka. Ubuduće je potrebno uvesti poboljšanja koja uključuju korištenje sustava za ekstrakciju imenovanih entiteta, ali je potrebno u oblikovanje same metode uvesti i informacije dobivene od sintaktičkog parsera.

Idući i završni korak oblikovanja primjera za učenje je odabir parova spominjanja. U ovom je sustavu to izvedeno uz korištenje tri ranije opisane heuristike (v. poglavlje 2). Dobiveni rezultati dobiveni su vrednovanjem uz pomoć koreferentnih parova iz ranije opisanog korpusa za hrvatski jezik (v. poglavlje 4). Metoda postiže niskih 17.2% odziva i 1.4% preciznosti koristeći automatski ekstrahirana spominjanja, gledajući broj korektno oblikovanih parova za učenje. Ukoliko se pak koriste isključivo ručno označena spominjanja, odziv raste na 68.1%. Ovaj ostatak neotkrivenih parova



Slika 7.1: Optimizacija parametra C

rezultat je nesavršenosti heurističkog određivanja roda i broja spominjanja pa brojni parovi ne prođu taj filtar.

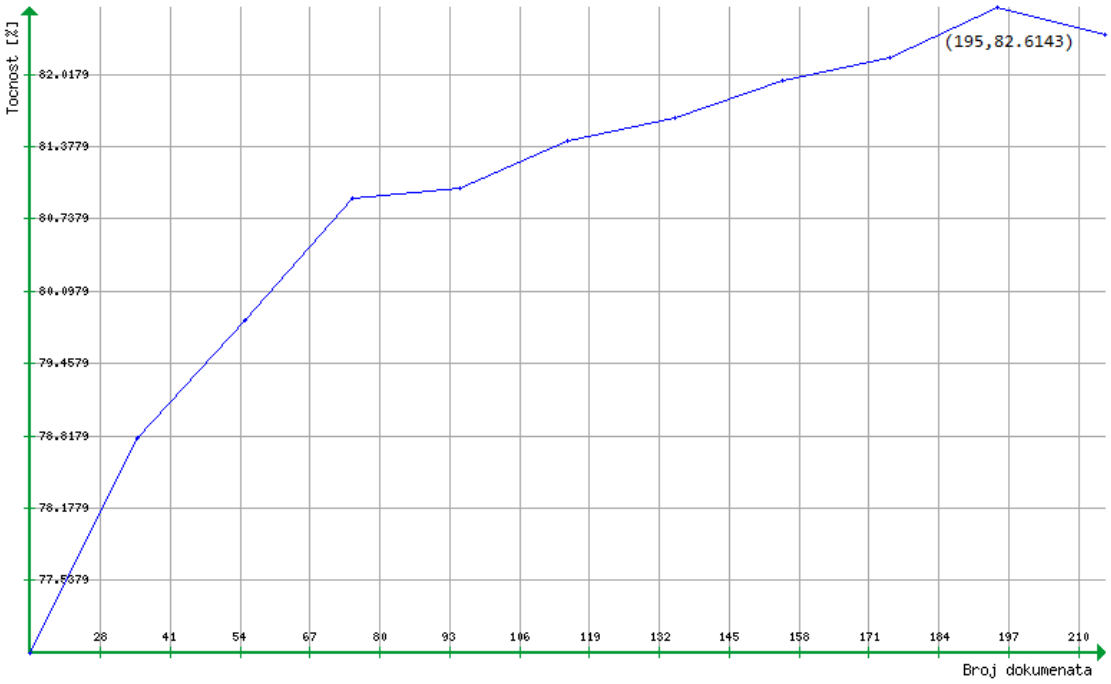
Ovako loši rezultati oblikovanja primjera za učenje doveli su do odluke da se daljnji eksprimenti provedu koristeći ručno označena spominjanja i odgovarajuće koreferentne parove iz korpusa.

7.2.2. Optimizacija parametra C

Provedena je optimizacija parametra C za stroj s potpornim vektorima. Optimizacija je izvedena treniranjem modela uz ispitivanje vrijednosti parametra C od 2^{-5} do 2^{20} koristeći peterostruku unakrsnu provjeru. Rezultati pretrage vidljivi su na slici 7.1. Vidljivo je da je najveća točnost postignuta uz vrijednost parametra $C = 2^{14}$.

7.2.3. Ovisnost rezultata o veličini skupa za učenje

Od ukupno 265 dokumenata, 50 je dokumenata rezervirano za testiranje, dok je preostalih 215 dokumenata bilo na raspolaganju za treniranje. Rezultati vrednovanja prikazani su na slici 7.2. Vidljivo je da se najbolji rezultati postižu s već oko 195 dokumenata, što govori da uz trenutnu izvedbu postupka nema smisla označavati još dokumenata.



Slika 7.2: Ovisnost o veličini skupa za učenje

7.2.4. Vrednovanje postupka evaluatorom

Nakon optimizacije vrijednosti parametra C i uz treniranje na svih 215 dokumenata, dobiveni je model testiran na skupu od 50 dokumenata. Dobiveni rezultati vrednovanja evaluatorom prikazani su u tablici 7.1. S obzirom na korištenje ručno označenih spominjanja, rezultati mjerenja točnosti otkrivanja spominjanja nisu pretjerano zanimljivi. Ono što je vrijedno pažnje su iznimno zadovoljavajući rezultati grupiranja spominjanja.

Ukupna službena mjera zadatka CoNLL iznosi 73.9% F1-mjere, što je daleko iznad najboljih rezultata za engleski jezik koji iznose 61.4% s također ručno označenim spominjanjima (Lee et al., 2011). Jasno je da ti rezultati nisu usporedivi zbog razlika u jeziku i korpusu, ali gledajući i druge jezike, ovo su vrlo obećavajući rezultati.

7.2.5. Vrednovanje skupova značajki

Provedeno je i vrednovanje odabira podskupova značajki metodom ablacije. U tablici 7.2 vidljivi su rezultati ispitivanja modela treniranih na svih 215 dokumenata i testiranih na 50 dokumenata. Najbolje rezultate očekivano daje korištenje svih dostupnih značajki, ukupno 82.4% točnosti klasifikacije parova.

Provedeni su eksperimenti s do jedne, dvije i tri pojavnice po spominjanju, kako

Tablica 7.1: Vrednovanje postupka evaluatorom

	Otkrivanje spominjanja			Grupiranje spominjanja		
Mjera	P	R	F1	P	R	F1
MUC	81.4	93.2	86.9	75.6	94.3	83.9
B-CUBED	81.4	93.2	86.9	66.2	96.0	78.3
CEAFe	81.4	93.2	86.9	77.6	48.1	59.4
BLANC	81.4	93.2	86.9	63.9	89.4	66.6
Ukupno	81.4	93.2	86.9	73.1	79.47	73.9

Tablica 7.2: Vrednovanje skupova značajki

Podskup	Točnost
Do 3 pojavnice	80.6%
Do 2 pojavnice	77.2%
Do 1 pojavnice	75.9%
Bez sint	80.05%
Bez MED	73.8%
Bez r/b	80.9%
Bez MED, r/b	73.6%
Sve	82.4%

bi se vidio utjecaj korištenja većeg broja pojavnica po spominjanju. Vidljivo je da se rezultati značajno poboljšavaju kako se upotrebljava sve više pojavnica po spominjanju. Maksimalno je korišteno do četiri pojavnice po spominjanju, što daje najbolje rezultate.

Zatim je ispitan utjecaj dostupnih značajki na razini para spominjanja, rod i broj (r/b) te Levenshteinova udaljenost između dvaju spominjanja (MED). Vidljivo je da su obje značajke korisne, s time da je Levenshteinova udaljenost iznimno korisna jer njezino izostavljanje povlači pad od 8.6 postotnih bodova.

Provjeren je još i utjecaj sintaktičkih značajki na rezultate. Kako je ranije opisano, vrsta sintaktičke veze navedena je za svaku pojavnicu u spominjanju. Izbacivanjem tih značajki točnost pada na 80.05%. Sintaktičke značajke imaju veći značaj pri ekstrakciji samih spominjanja, dok je njihov utjecaj na grupiranje spominjanja nešto manji, ali ipak značajan.

8. Zaključak

Razrješavanje koreferencije u tekstu jest problem otkrivanja i ekstrakcije spominjanja entiteta u tekstu te njihovo grupiranje po entitetu iz stvarnog svijeta. Ovo je sam po sebi vrlo težak problem, ali je njegovo rješavanje preduvjet za razvoj i poboljšanje naprednijih sustava za strojno prevođenje, automatsko odgovaranje na pitanja ili sažimanje teksta.

U ovom je radu opisan pristup razrješavanju koreferencije u tekstovima na hrvatskome jeziku, temeljen na metodama nadziranog statističkog strojnog učenja u kombinaciji s metodama temeljenim na ručno oblikovanim pravilima. Pristup je oblikovan proučavanjem trenutno najboljih sustava za engleski jezik, ali i dostupnih sustava u srodnim jezicima, poput češkog, poljskog i bugarskog.

Dobiveni rezultati ukazuju na prijeko potrebna poboljšanja metode oblikovanja primjera za učenje, dok suprotno tome metoda za grupiranje spominjanja pokazuje rezultate iznad očekivanja. Ukupni rezultat grupiranja spominjanja koristeći službeni evaluator s CoNLL 2011 iznosi 73.9% F1-mjere.

Budući rad bi se trebao sastojati s jedne strane od označavanja veće količine podataka, a s druge strane u poboljšavanju raznih komponenata postupka. S obzirom da novu dostupnost raznih naprednijih jezičnih alata za hrvatski jezik, trebalo bi bolje uklopiti informacije iz ovisnosnog parsera te iskoristiti ekstraktor imenovanih entiteta koji je također na raspolaganju.

To znači potpunu preradu metode za otkrivanje i ekstrakciju kandidata spominjanja iz teksta, ali i poboljšati selekciju parova spominjanja za učenje. Također bi trebalo uvesti naprednije metode grupiranja spominjanja koristeći poznate metode nenadziranog strojnog učenja.

LITERATURA

Ž. Agić i D. Merkler. Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. U *Text, Speech and Dialogue. Lecture Notes in Computer Science. Springer, in press*, 2013.

Ž. Agić, N. Ljubešić, i D. Merkler. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. U *Proceedings of BSNLP 2013, in press*, 2013.

C. Aone i S. Bennett. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. U *ACL*, stranice 122–129, 1995.

A. Bagga i B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. U *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, stranice 79–85, 1998.

R. Bellman i S. E. Dreyfus. *Applied Dynamic Programming*. 1962.

B. E. Boser, I. M. Guyon, i V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. U *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, stranice 144–152, 1992.

B. Broda, M. Marcińczuk, M. Maziarz, A. Radziszewski, i A. Wardyński. KPWr: Towards a Free Corpus of Polish. U *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

C.-C. Chang i C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology, Vol. 2, Issue 3*, stranice 27:1–27:27, 2011.

C. Cortes i V. N. Vapnik. Support-Vector Networks. U *Machine Learning*, stranice 273–297, 1995.

R. Florian, J. F. Pitrelli, S. Roukos, i I. Zitouni. Improving Mention Detection Robustness to Noisy Input. U *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, stranice 335–345, 2010.

G. Glavaš, M. Karan, F. Šarić, J. Šnajder, J. Mijić, A. Šilić, i B. Dalbelo Bašić. Croner: A State-of-the-Art Named Entity Recognition and Classification for Croatian. U *Proceedings of the Eighth Language Technologies Conference*, stranice 73–78, 2012.

A. Haghighi i D. Klein. Coreference Resolution in a Modular, Entity-Centered Model. U *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, stranice 385–393, 2010.

I. Hendrickx i V. Hoste. Coreference Resolution on Blogs and Commented News. U *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications*, stranice 43–53, 2009.

Z. Hranj. Razrješavanje koreferencije metodom nenadziranog strojnog učenja, 2011.

C.-W. Hsu, Chang C.-C., i Lin C.-J. A Practical Guide to Support Vector Classification, 2010.

P. Jain, M. R. Mital, S. Kumar, A. Mukerjee, i A. M. Raina. Anaphora Resolution in Multi-Person Dialogues, 1998.

I. Kmetovic. Uparivanje koreferentnih imenovanih entiteta metodama strojnog učenja, 2011.

J. K. Kummerfeld, M. Bansal, D. Burkett, i D. Klein. Mention Detection: Heuristics for the OntoNotes Annotations. U *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, stranice 102–106, 2011.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, i D. Jurafsky. Stanford's Multi-pass Sieve Coreference Resolution System at the Conll-2011 Shared Task. U *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, stranice 28–34, 2011.

V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady, Vol. 10, Num 8*, stranice 707–710, 1966.

X. Luo. On Coreference Resolution Performance Metrics. U *In Proceedings of HLT/EMNLP*, stranice 25–32, 2005.

J. F. McCarthy i W. G. Lehnert. Using Decision Trees for Coreference Resolution. U *Proceedings of the 14th International Joint Conference on Artificial intelligence - Volume 2*, stranice 1050–1055, 1995.

A. Nedoluzhko, J. Mirovsky, R. Ocelak, i J. Pergler. Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. U *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquim*, 2009.

V. Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. U *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, stranice 1396–1411, 2010.

V. Ng i C. Cardie. Improving Machine Learning Approaches to Coreference Resolution. U *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, stranice 104–111, 2002.

M. Ogrodniczuk, M. Zawislawska, K. Glowinska, i A. Savary. Coreference Annotation Schema for an Inflectional Language. U *CICLing (1)*, stranice 394–407, 2013.

S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, i N. Xue. Conll-2011 Shared Task: Modeling Unrestricted Coreference in Ontonotes. U *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, stranice 1–27, 2011.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, i C. Manning. A Multi-Pass Sieve for Coreference Resolution. U *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, stranice 492–501, 2010.

W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, Vol. 66, Num. 336*, stranice 846–850, 1971.

M. Recasens i E. H. Hovy. BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering, Vol. 17, Num. 4*, stranice 485–510, 2011.

M. Recasens, M. A. Marti, i M. Taule. Where Anaphora and Coreference Meet. Annotation in the Spanish CESS-ECE Corpus. U *Proceedings of the RANLP*, 2007.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, i Y. Versley. Semeval-2010 Task 1: Coreference Resolution in Multiple Languages, 2010.

W. M. Soon, H. T. Ng, i D. C. Y. Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics, Vol. 7, Num. 4*, stranice 521–544, 2001.

A. Soraluze, O. Arregi, X. Arregi, K. Ceberio, i A. D. de Ilarraza. Mention Detection: First Steps in the Development of a Basque Coreference Resolution System. U *Proceedings of KONVENS 2012*, stranica 128.

M. Strube i C. Müller. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. U *IN PROCEEDINGS OF THE 41ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, stranice 168–175, 2003.

M. Strube, S. Rapp, i C. Müller. The Influence of Minimum Edit Distance on Reference Resolution. U *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, stranice 312–319, 2002.

K. Van Deemter i R. Kibble. On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics, Vol. 26, Num. 4*, stranice 629–637, 2000.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, i L. Hirschman. A Model-Theoretic Coreference Scoring Scheme. U *Proceedings of the 6th conference on Message understanding*, stranice 45–52, 1995.

Razrješavanje koreferencije u tekstovima na hrvatskome jeziku

Sažetak

Razrješavanje koreferencije postupak je kojim se utvrđuje koji se izrazi u tekstu dokumenta odnose na isti izvanjezični entitet. Koreferentni izrazi mogu biti vlastita imena, imeničke fraze ili zamjenice. Razrješavanje koreferencije važan je zadatak u okviru obrade prirodnog jezika te nužan preduvjet za mnoge zadatke ekstrakcije informacije. Radi se o izrazito semantičkom problemu koji je težak kako za označavanje podataka, tako i za automatizirano rješavanje i vrednovanje.

U okviru ovog rada proučeni su postupci i sustavi za razrješavanje koreferencije u tekstu. Razrađen je postupak za otkrivanje referentnih spominjanja i razrješavanje koreferencije u tekstovima na hrvatskome jeziku. Postupak se temelji na metodama strojnog učenja te kombinira klasifikaciju parova spominjanja i grupiranje referentnih spominjanja. Razvijena je programska implementacija postupka i primjenjena na označenom novinskom korpusu tekstova na hrvatskome jeziku. Provedeno je eksperimentalno vrednovanje točnosti ekstrakcije, analiza značajki i detaljna analiza pogrešaka. Ukupni rezultat grupiranja spominjanja koristeći službeni evaluator s CoNLL 2011 iznosi 73.9% F1-mjere.

Ključne riječi: obrada prirodnog jezika, ekstrakcija informacija, strojno učenje, razrješavanje koreferencije, stroj s potpornim vektorima

Coreference Resolution in Croatian Texts

Abstract

Coreference resolution is a process of determining which expressions in a textual document refer to the same real-world entity. Corefering expressions can be names, noun phrases, or pronouns. Coreference resolution is an important task in scope of Natural Language Processing and a necessary step in solving many Information Extraction tasks. It is a semantically difficult problem that is both difficult for annotation as it is for automatic solving and evaluation.

In the scope of this paper different methods and systems for coreference resolution in text were studied. A method for extracting mentions and coreference resolution in Croatian texts was developed. The method is based on a supervised machine learning model and it combines mention-pair classification and clustering of corefering mentions. A software implementation was developed and applied on an annotated newspaper corpus in Croatian. Experimental evaluation of extraction accuracy, feature analysis, and a detailed error analysis were conducted. The final coreference resolution evaluation result is 73.9% F1, using the official CoNLL 2011 scorer.

Keywords: Natural Language Processing, Information Extraction, Machine Learning, Coreference Resolution, Support Vector Machine