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

Zaštićeno licencijom
Creative Commons Imenovanje-Nekomercijalno-Bez prerada 3.0 Hrvatska
https://creativecommons.org/licenses/by-nc-nd/3.0/hr/

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3806

Ekstrakcija navoda iz novinskih objava na hrvatskome jeziku

Zoran Medić

Zagreb, lipanj 2014.

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

Zagreb, 13. ožujka 2014.

ZAVRŠNI ZADATAK br. 3806

Pristupnik: **Zoran Medić**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Ekstrakcija navoda iz novinskih objava na hrvatskome jeziku**

Opis zadatka:

Novinske su objave primaran izvor informacija o događajima, stoga sustavi ekstrakciju informacija iz novinskih tekstova pobuđuju veliko zanimanje. Mnogo je informacija u novinskim objavama izraženo u obliku navoda. Analiza takvih navoda ima niz primjena, od analize događaja do analize medija i komunikološke analize. U literaturi je predloženo nekoliko postupaka za automatsku ekstrakciju navoda i njihovih izvora.

U okviru završnoga rada potrebno je proučiti postupke za ekstrakciju navoda iz tekstova, uključivo postupke temeljene na pravilima i postupke temeljene na strojnom učenju. Razraditi postupak za ekstrakciju navoda iz novinskih objava na hrvatskome jeziku koji će ekstrahirati navode izražene u obliku upravnoga govora te ih povezivati s imenovanim entitetima kao izvorima navoda. Izgraditi prikladnu zbirku novinskih tekstova s ručno označenim navodima i njihovim izvorima. Implementirati sustav za ekstrakciju i pregledan prikaz navoda i njihovih izvora. Provesti iscrpno vrednovanje sustava na ispitnoj zbirci te detaljnu analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i programsku dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 14. ožujka 2014.
Rok za predaju rada: 13. lipnja 2014.

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

U današnjem svijetu količina dostupnih informacija u novinskim objavama iznimno je velika, zbog čega je samim korisnicima pretraživanje takvih objava u potrazi za traženim informacijama znatno otežano. Postavlja se pitanje kako se nositi s velikim brojem dostupnih informacija ako nas zanima samo mali dio od tog broja? Jedno od mogućih rješenja je postupak ekstrakcije informacija kojim se iz teksta izdvaja samo traženi dio informacija.

Kada govorimo o informacijama o događajima, kao primaran izvor nameću se novinske objave. Velik broj korisnika pretražuje novinske objave u svrhu pronalaska informacija o događajima u koje su uključene i informacije o navodima i izvorima istih. Zbog toga i sustavi za ekstrakciju navoda pobuđuju veliko zanimanje, kako kod velikog broja svakodnevnih korisnika interneta koji žele na jednostavan način doći do tražene informacije, tako i kod specijaliziranih korisnika koji se bave različitim analizama, od analize događaja pa do analize medija i komunikološke analize.

Sama ekstrakcija informacija definira se kao postupak kojim se iz teksta izvlače već postojeće specifične informacije sadržane u tekstu ili nove informacije koje nisu direktno sadržane u tekstu. Dva su generalna pristupa ekstrakciji informacija iz teksta: metode temeljene na pravilima i metode strojnog učenja. Metode temeljene na pravilima temelje se na ručno definiranim uzorcima koje tražimo i pravilima po kojima ih pronalazimo. S druge strane, metode strojnog učenja temelje se na statističkom postupku kojim se metode pomoću prethodno označenog skupa dokumenata uče pronalaziti željene uzorke. Drugačije rečeno, postupkom nadziranog učenja model sam stvara pravila za prepoznavanje uzoraka, dok ih kod simboličkih tehnika stvara čovjek.

U domeni ekstrakcije navoda iz novinskih objava, problemu se pristupa različito, prvenstveno s obzirom na oblike navoda koji se žele ekstrahirati. Navodi u obliku upravnog govora najčešće se izriču u ograničenom broju oblika, pa je slijedom toga tehnika ekstrakcije informacija pomoću pravila primjerenija ekstrakciji takvih navoda. Za razliku od upravnog govora, neupravni govor se može izreći na mnogo više načina, a i samu granicu između dijela rečenice koji predstavlja neupravni govor i ostatka rečenice često je teško odrediti. Zbog toga su za ekstrakciju takvih navoda primjerenije metode strojnog učenja.

U okviru završnog rada proučeni su postupci za ekstrakciju navoda iz tekstova, preciznije postupci ekstrakcije navoda temeljeni na pravilima i postupci temeljeni na strojnom učenju. Implementiran je sustav za ekstrakciju navoda i njihovih izvora temeljen na pravilima, koji iz novinskih objava ekstrahira navode u obliku upravnog govora. Sustav se temelji na pretraživanju teksta u svrhu pronalaska navoda s pripadajućim imenovanim entitetom, te izjavnim glagolom koji ih povezuje. Dodatno, razrađen je i postupak kojim se navodi povezuju s temama o kojima se u njima govori. Za taj postupak korištene su specifične informacije sadržane u internetskim stranicama novinskih tekstova. Izgrađen je i opisan sustav za prikaz navoda i njihovih izvora pomoću kojeg je moguće pretraživati zbirku navoda, imenovanih entiteta i tema o kojima navodi govore.

U nastavku rada dan je pregled proučenih metoda za ekstrakciju navoda, uključivo metoda temeljenih na pravilima i metoda temeljenih na strojnom učenju. Nakon opisa metoda, slijedi opis implementiranog modela, zajedno s rezultatima provedenog vrednovanja sustava. Vrednovanje sustava provedeno je nad prikladnom zbirkom novinskih objava s ručno označenim navodima i njihovim izvorima. Provedeno vrednovanje rezultiralo je brojnim prijedlozima za poboljšanje i proširenje sustava, te su iste navedene u nastavku rada.

2. Postupci za ekstrakciju navoda i njihovih izvora

Postupci za ekstrakciju navoda i njihovih izvora obuhvaćaju metode temeljene na pravilima i metode strojnog učenja.

Metode temeljene na pravilima podrazumijevaju prethodno proučavanje domene problema. Zbog toga je potrebno istražiti sve oblike uzoraka u kojima se pojavljuju navodi u novinskim objavama i prilagoditi ekstrakciju tim uzorcima. Prednost ove metode jest to što ne zahtijeva nikakve prethodno označene dokumente za postupak ekstrakcije.

S druge strane, metode strojnog učenja ne zahtijevaju pretjerano znanje o domeni problema, već same pronalaze uzorke koje želimo ekstrahirati. Ipak, da bi to uspješno obavljale, potreban je skup ručno označenih dokumenata pomoću kojih model pronalazi tražene uzorke i uči pravila za njihovo prepoznavanje.

U nastavku su dani pregledi obiju metoda opisanih u nekoliko radova iz domene ekstrakcije navoda iz novinskih objava.

2.1. Metode temeljene na pravilima

Metode temeljene na pravilima u postupku ekstrakcije navoda iz novinskih objava temeljene su prvenstveno na regularnim izrazima koji opisuju moguću konstrukciju teksta u kojem se može pojaviti navod i njemu pripadajući imenovani entitet. Regularni izrazi grade se na razini pojavnica (eng. *token*), a u njima su obuhvaćeni različiti elementi rečenice dobiveni njezinom analizom - od interpunkcijskih znakova do uloga riječi u rečenici.

Primjer jednog takvog regularnog izraza dan je u nastavku. Dani regularni izraz opisuje pojavu navoda na početku rečenice:

(1) navodnici NAVOD navodnici [,] glagol [prilog] [apozicija] entitet

Postupak ekstrakcije navoda korištenjem spomenutih regularnih izraza prilično je jednostavan. Ako regularni izrazi opisuju traženu konstrukciju rečenice, metoda pro-

lazi kroz sve rečenice u tekstu i za svaku provjerava podudara li se s nekim od regularnih izraza. Izrazi mogu opisivati i konstrukciju paragrafa ili čak cijelog dokumenta, čime se mijenja jedinica teksta po kojoj se iterira. Ukoliko se jedinica teksta podudara s nekim od regularnih izraza, iz nje se ekstrahira navod i imenovani entitet povezan s tim navodom.

U literaturi je opisano nekoliko postupaka za ekstrakciju navoda temeljenih na pravilima. Pouliquen et al. (2007) izradili su sustav naziva "NewsExplorer" koji koristeći pravila za ekstrakciju navoda i imenovanih entiteta koji su izrekli navod izgrađuje bazu istih. Sustav obrađuje novinske članke, pronalazi dijelove teksta koji zadovoljavaju definirane izraze, te iz njih ekstrahira navode i odgovarajuće entitete. U svom sustavu koriste i jednostavnu metodu za razrješavanje anafore u kojoj se entiteti povezuju unutar istog dokumenta, ali samo u jednostavnim slučajevima u kojima je dio imena spomenut nakon punog imena i prezimena govornika. Ipak, zbog jednostavnosti samog postupka povezivanja govornika s navodom, u kojem se s navodom povezuje navodu najbliži imenovani entitet, a ne subjekt, česte su pojave pogrešnog ekstrahiranja govornika.

Sustav je zanimljiv zbog činjenice da se primjenjuje na nekoliko jezika, budući da je dio pravila pisan kako bi funkcionirao za različite svjetske jezike. Unutar sustava implementirana su tri generička pravila primjenjiva na više jezika, te nekoliko pravila specifičnih za pojedine jezike. Tri korištena generička pravila su:

1. *navodnici* NAVOD *navodnici* [,] *glagol* [*titula*] entitet.

   primjer: *"Sazvat ću sjednicu u subotu 7. lipnja", rekao je predsjednik Hrvatskog sabora Josip Leko.*

2. entitet [*, tekst duljine do 60 znakova,*] *glagol* [:lda] *navodnici* NAVOD *navodnici*

   primjer: *Josip Leko, predsjednik Hrvatskog sabora, izjavio je: "Sazvat ću sjednicu u subotu 7. lipnja".*

3. *navodnici* NAVOD *navodnici* [;l,] [*titula*] entitet *glagol*

   primjer: *"Sazvat ću sjednicu u subotu 7. lipnja", predsjednik Leko je dodao.*

Sustav je ostvario preciznost od 87.5% za engleski jezik, te odziv od 54%. Razlozi niskog odziva mogu se pronaći u činjenici da su pravila prilično strogo definirana, Primjerice, pojava jednog priloga između navoda i izjavnog glagola automatski onemogućuje ekstrakciju tog navoda. S druge strane, pogreške koje su utjecale na preciznost najčešće su bile one u kojima bi se govornikom proglasio krivi entitet. Naime, da bi entitet bio proglašen govornikom nije potrebno da bude subjekt, čime se posvojni pridjevi nepravilno proglašavaju govornicima (npr. *Milanovićev glasnogovornik*).

De La Clergerie et al. (2011) u svom su radu opisali izgradnju sustava “Sapiens” koji također obrađuje novinske tekstove u svrhu ekstrakcije navoda pomoću regularnih izraza izgrađenih na razini rečenice. Za razliku od “NewsExplorer”-a, “Sapiens” ekstrahira i slučajeve miješanog upravnog i neupravnog govora. Primjer jedne takve rečenice jest:

(2) *"To je za mene velika odgovornost i obveza. Posla se nikada nisam bojao, a u idućih pet godina puno se toga može napraviti za Istru i za sve građane Hrvatske", rekao je Jakovčić na konferenciji za novinare, dodajući kako "svojim znanjem i iskustvom može puno toga napraviti za sve one županije, općine i gradove" koji budu imali interes surađivati s njim.*

U drugom dijelu navedene rečenice dio unutar navodnika predstavlja primjer djelomičnog upravnog govora, budući da se u tom dijelu doslovno prenose riječi govornika.

Sustav koji su implementirali razrješava i anaforu u obliku zamjenica koje se referenciraju na prethodno navedene imenovane entitete. U postupku razrješavanja anafore vode se jednostavnim pravilom u kojem se kao mogući izvori navoda razmatraju entiteti spomenuti u dijelu teksta prije navoda. U prvom se koraku eliminiraju oni entiteti koji ne odgovaraju rodu i broju zamjenice, dok se u drugom koraku preostali entiteti rangiraju temeljem definiranih značajki. Korištene značajke uključuju gramatičku funkciju entiteta u rečenici, udaljenost entiteta od zamjenice, broj pojavljivanja entiteta u dijelu teksta koji prethodi zamjenici i pojavu entiteta unutar samog navoda.

Krestel et al. (2008) implementirali su u sklopu svog rada posebnu komponentu sustava “GATE” koja iz novinskih članaka ekstrahira navode u obliku upravnog, ali i neupravnog govora. Spomenuti sustav “GATE” je radni okvir koji podržava razvoj različitih aplikacija za obradu prirodnog jezika (Cunningham et al., 2002). Sustav je temeljen na pravilima koja se razlikuju prema položaju entiteta, glagola kojim se izriče navod i navoda u rečenici. Temeljna značajka sustava jest skup glagola kojima se izriče upravni ili neupravni govor (izjavni glagoli). Korištenjem navedenog skupa, u tekstu se najprije pronalaze pojave glagola iz tog skupa, te se ovisno o njihovoj poziciji u rečenici u odnosu na ostale elemente rečenice određene u pravilima za ekstrakciju (imenovani entiteti, navodnici) nastavlja obrada rečenice.

Vrednovanje sustava proveli su nad zbirkom ručno označenih navoda s ukupno 133 označena navoda. Pri tome su označili navod, govornika, izjavni glagol, ali i ostale dijelove rečenice koji nose dodatne informacije o navodu (engl. *circumstantial information*). Rezultati vrednovanja pokazali su da je preciznost sustava 98%, dok je odziv 83%.

2.2. Metode strojnog učenja

U postupku ekstrakcije navoda pomoću metoda strojnog učenja u literaturi je predloženo nekoliko metoda kojim su korištenjem različitih algoritama postignuti zadovoljavajući rezultati.

Elson i McKeown (2010) u svom su radu opisali postupak identifikacije navoda i govornika unutar književnih tekstova korištenjem metoda strojnog učenja. Naglasak je stavljen na povezivanje upravnog govora u dijalogu sa osobama koje su ga izrekle, točnije književnim likovima prisutnim u tekstu. Pri tom su kao glavnu značajku u povezivanju koristili informacije o entitetima koji su izrekli navode koji prethode onom koji se trenutno obrađuje. Iako je problem specifičan za književne tekstove, dobiveni rezultati pokazuju da se isti postupak može primijeniti i na različite tekstove koji ne sadržavaju nužno dijaloge u kojima govornici naizmjenično izriču navode, već i tekstove koji sadrže bilo kakav oblik upravnog govora povezan s govornikom. Postupak predložen u radu podijelili su u tri dijela: predobradu, klasifikaciju i učenje.

Predobrada podataka uključuje identifikaciju svih imenovanih entiteta i imenica koji prethode navodu koji se obrađuje. Identificirani eniteti i imenice predstavljaju potencijalne govornike za trenutni navod. Osim toga, obavljaju se i zamjene određenih dijelova teksta s jedinstvenim simbolima. Primjerice, izjavni glagol zamjenjuje se jedinstvenim simbolom "<EXPRESS_VERB>". Također, iz teksta se eliminiraju dijelovi koji ne sadrže informacije bitne za daljnju obradu (primjerice, rečenice i paragrafi u kojima nije naveden nijedan imenovani entitet).

Klasifikacija navoda podrazumijeva dodjeljivanje određene kategorije dijelovima teksta u kojima su prepoznati navodi. Klasifikacija se vrši korištenjem algoritma za podudaranje uzoraka, a korištene kategorije su:

- **Dodani navod** - Obuhvaća navode koji slijede prethodni navod u istom paragrafu.
- **Zasebni navod** - Obuhvaća navode koji se pojavljuju sami u paragrafu bez informacije o mogućim govornicima u istom paragrafu.
- **Trigram** - Predstavlja permutacije triju tokena: navoda, izjavnog glagola i govornika. Sastoji se od šest potkategorija, po jedna za svaku permutaciju. Primjerice, rečenica:'*"Pobjedili smo!" rekao je Ivan.*' pripada u kategoriju Navod-Glagol-Osoba.
- **Anafora trigram** - Kategorija slična kategoriji trigram, koja se također sastoji od šest permutacija različitih tokena. Razlika je u tokenu Osoba, koji je u ovom slučaju predstavljen zamjenicom, zbog čega kategorija i nosi ime "Anafora trigram".
- **Backoff** - Kategorija koja obuhvaća sve oblike navoda koji ne zadovoljavaju

nijednu od prethodno navedenih kategorija.

Učenje modela uključuje izgradnju vektora značajki za svaki par entiteta i navoda, te treniranje samog modela.

Značajke korištene za izgradnju vektora značajki za par navod-entitet obuhvaćaju:

- Udaljenost u riječima između entiteta i navoda
- Prisutnost bilo kojeg interpunkcijskog znaka između entiteta i navoda
- Redni broj entiteta između ostalih entiteta pronađenih u blizini navoda, sortiranih od onog najbližeg navodu do najdaljeg.
- Broj navoda prethodno izgovorenih od strane entiteta
- Broj entiteta, navoda i riječi u paragrafu
- Broj pojavljivanja entiteta u tekstu
- Za svaku riječ u blizini entiteta i navoda informacija je li riječ izjavni glagol, interpunkcijski znak ili drugi entitet
- Različite informacije o samom navodu, primjerice duljina navoda, pozicija unutar paragrafa, prisutnost entiteta unutar navoda, itd.

Nakon što su izgrađeni vektori značajki za svaki par entitet-navod, potrebno je od svih njih izabrati pravog govornika. U tu svrhu istraženo je nekoliko načina uspoređivanja vektora značajki pojedinog govornika s vektorima značajki preostalih govornika. Pritom su kao mjere usporedbe odabrane prosječne vrijednosti vektora, medijan vektora, minimalna i maksimalna vrijednost vektora i produkt značajki vektora. Uspoređivanje se obavlja tako da se za svaki izgrađeni vektor računa relativna udaljenost izgrađenog vektora od neke od predloženih mjera usporedbe. Ta udaljenost predstavlja ulaz u model strojnog učenja koji zatim odabire konačnog govornika.

U postupku učenja korištena su tri modela: C4.5 algoritam za izgradnju stabla odlučivanja, RIPPER algoritam, te model logističke regresije. Budući da ti modeli daju binarni izlaz za svakog mogućeg govornika posebno, posljednji korak je odabir najboljeg kandidata. Pri tom su isprobane četiri metode u kojima se govornik odabirao na temelju različitih kriterija. U "label" metodi, govornik se odabirao na temelju binarnog izlaza modela koji je govorio je li govornik traženi govornik ili ne. U metodi "jednostruke vjerojatnosti" govornik se odabire na temelju izlaza modela koji daje vjerojatnost da je govornik uistinu traženi govornik. "Hibridna" metoda funkcionira kao i metoda "label", s tim da se u slučaju više govornika označenih kao traženih koristi metoda "jednostruke vjerojatnosti" kako bi se odabrao jedan od njih. Metoda "kombinirane vjerojatnosti" funkcionira slično kao i metoda "jednostruke vjerojatnosti". Razlika je u tome što se za računanje vjerojatnosti koriste izlazi iz dva ili sva tri modela korištena za učenje.

Tako izgrađenim sustavom postignuti su rezultati preciznosti od 83%. Posebno dobri rezultati postignuti su u kategoriji "Dodani navod", u kojoj su klasifikacijski modeli izveli pravila slična onima koja su autori ručno izveli.

O'Keefe et al. (2012) su koristeći prethodno opisani rad implementirali sustav za ekstrakciju navoda korištenjem metoda za označavanje slijednih podataka (engl. *sequence labbeling*). Za klasifikaciju navoda i govornika koristili su dva klasifikatora: binarni i višeklasni klasifikator.

Binarni klasifikator za svaki je navod razmatrao n prethodno spomenutih entiteta i za svakog od njih posebno računao vjerojatnost povezanosti s zadanim navodom, pri čemu bi svaki od entiteta bio označen ili kao "govornik" ili kao "ne-govornik". Na kraju je od svih entiteta koji su označeni kao "govornici" izabran onaj s najvećom vjerojatnošću.

Višeklasni klasifikator za svaki je navod razmatrao svih n prethodno spomenutih entiteta, te pri računjanju vjerojatnosti za pojedini entitet uzimao u obzir i ostale entitete, odnosno koliko su "dobri" ostali kandidati. Na kraju bi, kao i binarni, izabrao onaj entitet s najvećom vjerojatnošću.

Glavna zamjerka radu Elson i McKeown (2010) bila im je korištenje informacije o govornicima prethodnih navoda kao značajke u modelu, budući da su pretpostavke o besprijekorno točnom povezivanju entiteta s navodom prilično nerealne. Zbog toga su u svom radu tom problemu pristupili upravo kao zadatku označavanja slijednih podataka, gdje je kao slijed koji je potrebno označiti definiran skup govornika u dokumentu. Tako je u svakom stanju, točnije navodu čiji se govornik traži, poznato w prethodno povezanih govornika s navodom, te odluka o govorniku za trenutni navod. Pri tome su koristili tri tehnike za efikasno učenje: pohlepni algoritam, Viterbijev algoritam i metodu nasumičnih uvjetnih polja (engl. *Conditional Random Fields - CRF*).

U pohlepnom algoritmu, u svakom se koraku primjenjuje standardni klasifikator u kojem se značajke za slijedno označavanje računaju iz rezultata (govornika) pretpostavljenih u prethodnim koracima. Ovakav je algoritam potpun budući da u svakom koraku razmatra samo jedan slijed prehodno definiranih odabira. Međutim, nije optimalan, jer nije sposoban prepoznati razlike između dobrih trenutnih i prethodnih odabira.

Viterbijev algoritam pronalazi najvjerojatniji slijed govornika pomoću niza odluka kojima povezuje govornika s navodom. U svakom se navodu, za svaki od mogućih govornika, računa vjerojatnost povezivanja govornika s navodom na temelju svih mogućih kombinacija povezivanja nekog od govornika s w prethodnih navoda. Pri tome se za svakog govornika odredi najvjerojatniji slijed, te se vjerojatnosti tog slijed pomnože s vjerojatnošću trenutnog govornika, te se kao konačan odabir uzima onaj s najvećom ukupnom vjerojatnošću. Glavna razlika u odnosu na pohlepni algoritam je u tome što se u Viterbijevom algoritmu uzimaju u obzir svi mogući prethodni slijedovi pri odabiru

trenutnog govornika, dok se u pohlepnom algoritmu uzima u obzir samo jedan takav slijed.

U metodi nasumičnih uvjetnih polja ulazni i izlazni skup, koje u ovom slučaju predstavaljaju skup govornika i navoda povezanih s govornicima, prikazan je u obliku grafa. U postupku treniranja modela, metoda je sposobna prepoznati veze između značajki koje opisuju vezu između govornika i određenog navoda i stvarne oznake govornika s navodom. Također, sposobna je i prepoznati vjerojatnost prijelaza iz jednog stanja koje opisuje označeni navod u drugo. Pomoću tih podataka, metoda obilazi graf i pronalazi ona stanja u kojima je vjerojatnost povezanosti nekog od govornika s određenim navodom najveća, te se tako konstruira potpuni skup navoda s povezanim entitetima.

Izgrađeni model evaluirali su na tri korpusa: dva sastavljena od novinskih objava i jednom sastaljenom od književnih tekstova. Na korpusima sastavljenim od novinskih objava postignuti su znatno bolji rezultati - preciznost je iznosila u prvom slučaju 92.4%, te u drugom 84.1%. Kod korpusa sastavljenog od književnih tekstova preciznost je bila znatno manja - 53.3%.

3. Model za ekstrakciju navoda

Implementirani model za ekstrakciju navoda temeljen je na postupku ekstrakcije navoda pomoću pravila. U tu svrhu izrađen je skup pravila za ekstrakciju, koji je sastavljen proučavanjem literature, gdje je kao glavna referenca korišten rad "Automatic Detection of Quotations in Multilingual News", (Pouliquen et al., 2007).

Iz tog sustava preuzeta su tri generička pravila za prepoznavanje navoda, te je dodano još nekoliko pravila kojim se ekstrahiraju uzorci koji nisu obuhvaćeni trima generičkim pravilima. Neka pravila su implementirana na općenitiji način, tako da obuhvaćaju više pravila definiranih u proučenim radovima.

3.1. Opis modela

Kao jedinica na čijoj razini se ekstrahiraju navodi odabrana je rečenica, budući da se navodi najčešće protežu kroz jednu rečenicu. Zbog toga se u postupku ekstrakcije navoda iterira kroz rečenice dokumenta dobivene predobradom teksta (rastavljanjem na rečenice).

Metoda korištena za ekstrakciju prilično je jednostavna. U obrađenom tekstu pokušavaju se pronaći uzorci koji odgovaraju definiranim pravilima. Pravila opisuju mjesta u rečenici gdje se može pojaviti navod, osoba koja je navod izrekla - subjekt, te glagol koji povezuje navod i osobu - predikat.

U svakom pravilu postavljeno je ograničenje na pojavnice koje predstavljaju subjekt koji je izrekao navod i predikat koji povezuje navod sa subjektom. Kako bi navod bio uspješno ekstrahiran potrebno je da subjekt bude imenovani entitet označen kao osoba, drugačije rečeno kao *"Person"* u izlazu iz alata CroNER. Predikat, točnije njegov lematizirani oblik - infinitiv, treba biti glagol iz skupa izjavnih glagola koji predstavljaju glagole kojima se izriče navod. Skup izjavnih glagola preuzet je iz literature, točnije iz rada Krestel et al. (2008) u kojem su navedeni korišteni izjavni glagoli. Preuzeti glagoli prilagođeni su za hrvatski jezik. Popis svih glagola iz korištenog skupa izjavnih glagola nalazi se u tablici 3.1.

Tri generička pravila izmijenjena su u implementiranom sustavu tako da je 1. i 3. pravilo spojeno u jedno novo pravilo u kojem se nakon završnih navodnika navoda s

argumentirati	biti	bojati	brinuti	citirati
dodati	dodavati	govoriti	gunđati	inzistirati
iskazati	ispričati	isticati	izgovoriti	izjaviti
izjavljivati	izvijestiti	izvještavati	kazati	komentirati
kritizirati	kukati	likovati	misliti	nadati
naglasiti	nagovijestiti	napisati	narediti	nastaviti
navesti	navoditi	obećati	obećavati	objasniti
objaviti	objašnjavati	obrazložiti	odati	odgovarati
odgovoriti	odlučiti	okriviti	opaziti	opisati
opisivati	opomenuti	opovrgnuti	optužiti	optuživati
osporavati	otkriti	otkrivati	pisati	pitati
podsjetiti	pojasniti	poreći	poručiti	poručivati
posvjedočiti	potvrditi	povjeriti	predložiti	predvidjeti
predviđati	preporučiti	preporučivati	pretpostaviti	primijetiti
priopćiti	prisjetiti	prisjećati	priznati	pričati
procijeniti	proglasiti	proreći	proricati	proturječiti
reći	savjetovati	sjetiti	sjećati	složiti
smatrati	smirivati	spomenuti	stajati	sugerirati
teretiti	tumačiti	tužiti	tvrditi	upitati
upozoriti	ustvrditi	uzvratiti	vjerovati	zahtijevati
zaključiti	zapovijediti	zaprijetiti	završiti	žaliti

Tablica 3.1: Skup izjavnih glagola

početka rečenice može pojaviti ili glagol ili entitet. U naknadnoj evaluaciji ustanovljeno je da su slučajevi s entitetom nakon navodnika zapravo dosta rijetki, pa je time objedinjavanje ovih dvaju pravila opravdano.

Osim navedenih ograničenja u pogledu subjekta i predikata rečenice, uvedeno je i ograničenje u pogledu teksta unutar navoda. Kako bi tekst unutar navoda bio prepoznat kao navod, potrebno je da započinje s velikim slovom. Na taj način sustav odbacuje dijelove teksta koji predstavljaju polu-upravni govor ili dijelove koji uopće ne predstavljaju upravni govor.

Sustav trenutno ne pokušava razriješiti moguću koreferenciju u rečenicama, pa tako entitet *Babić* iz primjera (3) neće biti povezan s eventualnom pojavom entiteta *Ivica Babić*, tj. punog imena i prezimena govornika, u ostatku teksta.

Također, sustav ne razrješava anaforu između zamjenica ili skrivenog subjekta (engl. *zero anaphora*) koji se odnose na entitet koji je izrekao navod i samog entiteta. Primjerice u rečenicama (3) i (4) koje se nalaze u istom članku, sustav ne povezuje

entitet iz rečenice (3) sa skrivenim subjektom iz rečenice (4), čime navod iz rečenice (4) ne ekstrahira.

(3) *"Ta odluka će ovisiti o kolektivnom ugovoru, koji još nije potpisan", rekao je Babić.*

(4) *"Nažalost, veliki problemi u zdravstvu pogađaju prije svega pacijente, ali i liječnike, posebno u bolnicama, preko čijih leđa se prelamaju brojni problemi. Rezultat katastrofalnog stanja u zdravstvu je sve češće najavljivanje odlaska liječnika u druge države", zaključio je.*

Osim takvog oblika anafore sustav ne ekstrahira ni oblike u kojima se govornik oslovljava imenicom povezanom s govornikom. Primjerice, sustav ne ekstrahira navod iz primjera (5), niti ako isti slijedi neposredno nakon rečenice iz primjera (3), budući da je za povezivanje entiteta *Babić* iz primjera (3) s izrazom *predsjednik Hrvatskog liječničkog sindikata* iz primjera (5) potreban alat za razrješavanje anafore, koji nije korišten u razvoju sustava.

(5) *"Nažalost, veliki problemi u zdravstvu pogađaju prije svega pacijente, ali i liječnike, posebno u bolnicama, preko čijih leđa se prelamaju brojni problemi. Rezultat katastrofalnog stanja u zdravstvu je sve češće najavljivanje odlaska liječnika u druge države", zaključio je predsjednik Hrvatskog liječničkog sindikata.*

3.2. Ekstrakcijska pravila

U skupu pravila za ekstrakciju nalaze se četiri pravila koja su dana u nastavku.

1. *navodnici* NAVOD *navodnici* [,] [*predikat*|subjekt] [riječi koje nisu subjekt ili predikat] [subjekt|*predikat*] [*].
2. [*pojavnice*] subjekt [*pojavnice koje nisu subjekt ni predikat*] *predikat* : *navodnici* NAVOD *navodnici*
3. *navodnici* NAVOD *navodnici* [,] [*] *predikat* [*] subjekt [*] *predikat* : *navodnici* NAVOD *navodnici*
4. - NAVOD , *predikat* [*] subjekt [*].

Prvo pravilo odnosi se na slučaj u kojem je navod na početku rečenice, nakon čega slijedi ili predikat iz skupa izjavnih glagola ili subjekt označen kao imenovani entitet i to osoba. Ako je neposredno nakon navoda naveden subjekt, u ostatku teksta pronalazi se predikat, dok se u slučaju pojave predikata neposredno uz navod u ostatku teksta pronalazi subjekt.

Drugo pravilo odnosi se na slučaj u kojem se navod nalazi na kraju rečenice. U prvom dijelu rečenice nalaze se subjekt i predikat, i to tim redoslijedom, a između predikata i početnih navodnika u navodu nalazi se dvotočje.

Treće pravilo odnosi se na slučaj u kojem se navod nalazi na dva mjesta u rečenici: na početku i na kraju rečenice. Između dva navoda nalaze se dva predikata i jedan subjekt. Prvi predikat povezuje subjekt sa prvim navodom, dok drugi predikat povezuje subjekt s drugim navodom. Kao i u ostalim pravilima, i u ovom predikati trebaju biti iz skupa izjavnih glagola, te subjekt mora biti imenovani entitet - osoba.

Četvrto pravilo odnosi se na poseban oblik izricanja navoda, koji se pokazao čestim u novinskim objavama. Ono obuhvaća slučaj u kojem je navod prvi dio rečenice, te počinje s oznakom crtice '-', a završava s posljednjim zarezom u rečenici iza kojeg slijedi predikat, a odmah nakon njega i subjekt. Pravilo obuhvaća samo slučajeve u kojima se tekst navoda sastoji od jedne rečenice, budući da bi se u slučaju navoda sastavljenog od više rečenica iste obrađivale odvojeno. Primjer rečenice koja zadovoljava objašnjeno pravilo jest:

(6) *- Ili odlazim nakon ove sezone, ili ću još jednu sezonu ostati u klubu u kojem sam sada, poručio je Balić.*

Prvi korak u analizi rečenice je potraga za navodnicima, budući da upravo oni označavaju pojavu upravnog govora. Nakon toga nastavlja se daljnja obrada ovisno o pravilu koje se obrađuje. Ta obrada uključuje potragu za predikatom - glagolom iz skupa izjavnih glagola, te subjektom - imenovanim entitetom označenim kao osoba. Po potrebi u rečenici se traže i ostale pojavnice (npr. zarezi ili dvotočje). U slučaju pojave rečenice koja se podudara s većim brojem pravila, odabire se prvo pravilo u redoslijedu koji je naveden u danom popisu pravila.

4. Vrednovanje sustava

Vrednovanje sustava provedeno je nad označenim skupom podataka koji se sastoji od ukupno 60 novinskih objava prikupljenih s navedenih portala. Ukupno je u tih 60 objava označeno 119 navoda s pripadajućim izvorima. Vrednovanje je pokazalo da je izgrađeni sustav ispunio očekivanja u domeni u kojoj pokušava ekstrahirati navode, točnije u uzorcima koji su pokriveni implementiranim pravilima. Ipak, postoje pogreške koje bi se daljnjim poboljšanjima sustava mogle ispraviti. Također, postoje i prijedlozi za proširenje sustava kojim bi se proširila domena unutar koje sustav pronalazi navode.

4.1. Ručno označena zbirka

Za potrebe vrednovanja sustava ručno je označeno 119 navoda u 60 novinskih članaka, što prosječno iznosi 1.98 navoda po članku. Pri označavanju označeni su samo navodi koji se pojavljuju unutar rečenice zajedno s entitetom koji je navod izrekao, budući da sustav samo takve navode i pokušava ekstrahirati.

U postupku označavanja u tekstu su označeni indeksi pojavnica koje označavaju početak i kraj navoda, te indeksi pojavnica koje označavaju početak i kraj imenovanog entiteta - govornika. Kako bi označivaču bili poznati ti podatci, tekst je prije označavanja rastavljen na rečenice, a zatim na pojavnice u rečenici, pri čemu je svakoj od njih pridružen indeks koji označuje početni indeks pojavnice u rečenici. Također, u rečenicama su označeni i imenovani entiteti, kako bi označivač znao je li govornik prepoznat kao osoba ili neki drugi tip entiteta.

Označavanje je obavila jedna osoba, budući da u samom postupku označavanja ne postoje neke značajne razmirice koje bi bilo potrebno usuglasiti. Kao početak i kraj navoda označeni su indeksi navodnika, dok su kao početak i kraj imenovanog entiteta označeni početna i završna pojavnica u izlazu iz označivača imenovanih entiteta. Koristeći takav postupak, moguće nesuglasice dane su na rješavanje alatima za predobradu podataka.

4.2. Rezultati vrednovanja sustava

Od 119 označenih navoda s pripadajućim govornicima, sustav je uspješno ekstrahirao njih 86 (*true positive*), dok preostala 33 navoda nije ekstrahirao (*false negative*). Također, sustav je ekstrahirao i 2 primjera koji su netočno kvalificirani kao pozitivni (*false positive*).

Ti rezultati doveli su do postotka preciznosti od 97.7%, što je usporedivo s rezultatima koja su u svom radu postigli Pouliquen et al. (2007), gdje je postignuta preciznost iznosila 87.5%. Razlog visoke preciznosti može se pronaći u činjenici da su pravila strogo definirana za specifične slučajeve, a i sama pojava uzoraka koji se podudaraju s definiranim pravilima znači najčešće upravo pojavu upravnog govora. Dobiveni odziv sustava iznosi 72.3%. Rezultat odziva također je usporediv s radom (Pouliquen et al., 2007) gdje je postignut odziv iznosio 64%. Odziv je u ovom slučaju manji od preciznosti, a razlog za to može se najvećim dijelom potražiti u alatima za predobradu podataka, budući da o njima dobrim dijelom ovisi i izlaz sustava. Korištenjem izračunatih podataka o preciznosti i odzivu, dolazi se do F1-mjere implementiranog sustava koja iznosi 83%.

4.3. Analiza pogrešaka

U izgrađenom sustavu postoji još prostora za poboljšanje i proširenje. Vrednovanje sustava ukazalo je na neke od pogrešaka koje su uglavnom nastale zbog pogreški generiranih tijekom predobrade teksta.

Tijekom vrednovanja sustav je ekstrahirao 2 primjera koji su netočno kvalificirani kao pozitivni. U oba slučaja, subjekt koji je izrekao ekstrahirani navod odnosio se na organizaciju, ali je netočno označen kao osoba. Primjer (7) predstavlja jedan od ta dva primjera.

(7) *"Bivši veznjak Tottenhama s nekoliko pametnih dodavanja pokazao je napadačku moć, no pravi posao je napravio u obrani gdje se stvarno isticao i zadržao Bayern", piše Eurosport koji je Modrića proglasio najboljim igračem madridskog spektakla.*

U navedenom primjeru entitet *Eurosport* koji označava organizaciju, pogrešno je označen kao osoba, čime je sami navod pogrešno okarakteriziran kao pozitivan primjer.

U vrednovanju je primjećeno da je kod velikog broja primjera koje je sustav netočno okarakterizirao kao negativne, razlog ležao u neispravnoj kategorizaciji subjekta, odnosno imenovanog entiteta. Primjerice, u sljedećoj rečenici entitet *Šojgu* označen je kao organizacija, a ne kao osoba, zbog čega navod nije uspješno ekstrahiran.

(8) *"Bili smo primorani reagirati na takav razvoj događaja", rekao je Šojgu.*

Taj tip pogreške primjećen je uglavnom u slučajevima u kojima je ime subjekta stranog podrijetla, pa bi se taj problem mogao riješiti i korištenjem označivača entiteta primjenjivog na više jezika.

Čest razlog pogreške bio je i poseban oblik izricanja upravnog govora u kojem se kao graničnici navoda (navodnici) koriste crtice (–). Takav oblik izricanja navoda nije prepoznat kao dio rečenice u korištenom alatu za razdvajanje rečenica. Samim time, rečenice koje sadrže takav oblik navoda bile bi rastavljene u više rečenica, naravno ukoliko se navod sastoji od više rečenica. Sustav zbog tog razdvajanja ne ekstrahira takve navode, budući da je jedinica na kojoj se navodi ekstrahiraju upravo rečenica, pa bi za ekstrakciju takvih navoda bilo potrebno formulirati pravila na razini paragrafa. Primjer jednog takvog navoda dan je u nastavku.

(9) *– Sve se to temelji na pretpostavci gospodarskog rasta. Država nam je skupa i neefikasna i to je kronična bolest u Hrvatskoj – rekao je premijer Zoran Milanović te je ispričao priču o uspjehu Švedske.*

Pogreške tog tipa mogle bi se riješiti na više načina. Jedan od mogućih je jednostavna zamjena crtice (–) sa uobičajenim dvostrukim navodnicima (“) i obrada tako dobivenog teksta. Međutim, takvim pristupom moguće su nove pogreške, budući da se crtica može pojaviti na više mjesta u rečenici, a ne samo kao oznaka početka ili kraja navoda. Drugi mogući pristup jest promjene razine na kojoj se navodi ekstrahiraju s razine rečenice na razinu paragrafa ili cijelog dokumenta. Takav pristup vjerojatno bi poboljšao i cjelokupne rezultate, jer je broj navoda koji se protežu kroz nekoliko paragrafa također značajan.

Nekoliko primjera nije uspješno ekstrahirano zbog neprepoznavanja subjekta u rečenici. Alat[1] korišten za gramatičku analizu rečenice nije uspješno prepoznao subjekt u rečenici, pa ni sami navod nije uspješno ekstrahiran, budući da u rečenici nije prepoznat govornik. Jedan takav primjer dan je u nastavku:

(10) *"Eksperiment je podsjetnik na to da su beskućnici ljudi, kao i mi, ali s jednom razlikom. Imaju problema i to ih boli. I da, oni mogu biti nečiji ujak, rođak ili supruga", rekao je Craig Mayes za HuffPost.*

U navedenom primjeru, entitet *Craig Mayes* označen je kao objekt, a ne subjekt, čime navod automatski nije ekstrahiran.

4.4. Prijedlozi za poboljšanje sustava

Iako je izgrađeni sustav pokazao dosta dobre rezultate gledajući preciznost i odziv sustava, postoji još prostora za poboljšanja i proširenja sustava, prvenstveno zbog toga

[1]MST Dependency Parser (prilagođen za hrvatski jezik) - nlp.ffzg.hr/resources/models/tagging/

što sustav trenutno obuhvaća samo domenu navoda u obliku upravnog govora.

Poboljšanja sustava moguća su prvenstveno različitim postupcima kojima bi se ispravile pogreške navedene u prethodnom poglavlju: neispravno označavanje entiteta, neispravno označavanje subjekta i neispravno razdvajanje rečenica. Osim rješavanja tih problema, sustav se može poboljšati i korištenjem novih alata kojima bi se proširila domena navoda izrečenih upravnim govorom koji se pokušavaju ekstrahirati. Također, moguće je i proširenje sustava na ekstrakciju neupravnog govora, budući da je isti znatno više zastupljen u novinskim objavama od upravnog govora.

Kao jedno od značajnijih proširenja sustava predlaže se korištenje alata za razrješavanje anafore. Korištenje tog alata značajno bi povećalo kvalitetu sustava, budući da se velik broj navoda u tekstu referencira na govornika preko imenice koja ga označava ili zamjenice. Nažalost, još nije dostupan takav alat za hrvatski jezik, pa sustav u ovom trenutku ne ekstrahira takve primjere.

Utjecaj na kvalitetu sustava imalo bi i proširenje domene navoda koje sustav ekstrahira na navode izrečene u neupravnom govoru. Ipak, ekstrakcija navoda u obliku neupravnog govora nije primjerena za metode temeljene na pravilima, iz razloga što se neupravni govor može izreći na mnogo više načina od upravnog govora. Uz to, granice neupravnog navoda često je teško odrediti, za razliku od upravnih navoda koji su od ostatka teksta odvojeni navodnicima. Zbog toga se za taj problem predlažu metode strojnog učenja koje bi obuhvatile širok spektar načina za izricanje neupravnog govora. Neke od tih metoda opisane su i u proučenim radovima (De La Clergerie et al., 2011; Krestel et al., 2008).

Jedno od mogućih proširenja sustava predstavlja i ekstrakcija navoda u kojima kao govornici nisu navedene isključivo osobe, već i organizacije. Takvo proširenje bilo bi posebno korisno pri ekstrakciji navoda iz novinskih objava iz poslovne domene, budući da se u istima često navode samo organizacije iz kojih su navodi potekli, a ne i osobe koje su ih izgovorile (najčešće glasnogovornici). Primjer takvog navoda dan je u nastavku.

(11) *”U sklopu ovog projekta u sniženje cijena uloženo je više od 100 milijuna kuna“, kažu iz Konzuma.*

5. Programska izvedba

Programska izvedba ostvarana je u programskom jeziku Java. Grafičko korisničko sučelje izrađeno je korištenjem tehnologije JSP - Java Servlet Pages, dok je za pohranu zbirke navoda i pripadajućih imenovanih entiteta korištena relacijska baza podataka MySQL.

Izgrađeni sustav može se podijeliti u nekoliko sastavnica:

1. Podsustav za prikupljanje novinskih objava s Interneta
2. Podsustav za predobradu podataka (eng. *preprocessing*)
3. Podsustav za ekstrakciju navoda iz obrađenog teksta
4. Podsustav za prikaz podataka na grafičkom korisničkom sučelju

Podsustav za prikupljanje novinskih objava s Interneta (eng. *crawler*) prikuplja novinske članke s nekoliko većih internetskih portala parsirajući HTML sadržaj stranice. Članci prikupljeni tim putem pohranjuju se kao tekstne datoteke spremne za daljnju obradu.

Podsustav za predobradu podataka vrši potrebnu predobradu podataka unutar teksta članka korištenjem različitih alata za analizu rečenične strukture. Alati vrše sljedeće postupke analize rečenice:

- Rastavljanje teksta na rečenice (eng. *sentence splitting*)
- Označavanje imenovanih entiteta unutar rečenice
- Rastavljanje rečenice na pojavnice (eng. *tokenization*)
- Lematizacija
- Gramatičko označavanje dijelova rečenice (eng. *part-of-speech tagging*)

Tako obrađen tekst prilagođen je daljnjoj obradi budući da se u njemu lako može pristupiti dijelovima rečenice bitnim za daljnji tijek obrade i ekstrakcije navoda.

Podsustav za ekstrakciju navoda iz obrađenog teksta temeljen je na pravilima za prepoznavanje navoda i njihovih izvora. Ukupno su implementirana četiri pravila s

kojima se uspoređuju dijelovi teksta i pokušavaju pronaći uzorci koji odgovaraju pravilima. Pri ekstrakciji provjeravaju se sve razine rečenične strukture, od najniže razine koja uključuje pojavnice do gramatičkih kategorija riječi u rečenici.

Podsustav za prikaz podataka na grafičkom korisničkom sučelju sastoji se od grafičkog korisničkog sučelja implementiranog u obliku jednostavne web-stranice i prikladne baze podataka u kojoj su pohranjeni navodi, njihovi izvori, adrese web-stranica s na kojima se nalaze originalni tekstovi, originali tekstovi članaka, te teme o kojima se u navodima govori. Korisnicima je omogućeno pretraživanje zbirke navoda prema izvorima navoda, ali i temama o kojima govore.

U nastavku je svaki od podsustava detaljno opisan, zajedno sa sastavnicama koje ga čine.

5.1. Podsustav za automatsko prikupljanje novinskih objava

Podsustav za automatsko prikupljanje tekstova novinskih objava izgrađen je kao web-klijent popularno zvan pauk (engl. *spider* ili *crawler*) koji obrađuje sljedeće web-portale:

- www.index.hr
- www.jutarnji.hr
- www.vecernji.hr

S navedenih portala prikupljaju se članci te se njihov sadržaj pohranjuje u tekstne datoteke. Tekstne datoteke u prvom retku imaju pohranjen link na članak, te popis tema o kojima se u članku govori. Nakon toga slijede paragrafi članka, svaki u svom retku.

Web-klijent je izveden u programskom jeziku Java, korištenjem javno dosupne biblioteke JSoup[1]. JSoup je biblioteka pisana u Javi koja pruža mogućnosti parsiranja HTML-a, kao i jednostavno sučelje za pronalaženje i ekstrakciju željenih podataka iz HTML zapisa.

Korištenjem JSoup biblioteke iz HTML stranice ekstrahira se tekst članka, ali i dijelovi stranice koji se sastoje od ključnih riječi koje opisuju teme o kojima se u članku govori. Te riječi koriste se u nastavku obrade kao teme navoda iz pripadajućeg članka, te se spremaju u bazu podataka skupa s navodima iz tog članka. Iako ne govore svi navodi o svakoj temi opisanoj nekom od ključnih riječi, pristup se pokazao korisnim budući da se ključne riječi najčešće birane tako da opisuju općenite teme članka. U

[1] JSoup biblioteka - jsoup.org

budućnosti bi se za kategorizaciju navoda mogao koristiti neki od sustava za kategorizaciju teksta, čime bi se poboljšala preciznost kategorizacije samih navoda.

5.2. Podsustav za predobradu podataka

Podsustav za predobradu podataka sastoji se kao što je navedeno od nekoliko sastavnica od kojih je svaka obavlja jedan od postupaka analize rečenice.

U nastavku je detaljnije opisana svaka od sastavnica.

Rastavljanje teksta na rečenice

Rastavljanje teksta na rečenice obavljano je korištenjem prikladnih alata za hrvatski jezik. Budući da je tekst članka u datotekama spremljen tako da je u svakom retku zapisan jedan paragraf, rastavljanje teksta na rečenice znači rastavljanje svakog paragrafa na pripadne rečenice.

Označavanje imenovanih entiteta

Označavanje imenovanih entiteta unutar rečenice obavljeno je korištenjem alata CroNER (Glavaš et al.). Alat CroNER[2] unutar danog teksta prepoznaje imenovane entitete i razvrstava ih u nekoliko kategorija. Za problem povezivanja izvora navoda sa samim navodom korisna je kategorija "Person" kojom CroNER označava svaku pojavu imenovanog entiteta koja označava osobu.

Rastavljanje rečenice na pojavnice

Rastavljanje rečenice na pojavnice ostvareno je korištenjem prikladnog alata za hrvatski jezik. Pojavnice predstavljaju sve ono što se nalazi između dva pismena koja služe kao graničnici, pri čemu su ona pismena koja se nalaze između graničnika simboli abecede kojima su pridodane znamenke i crtica. Tako rastavljena rečenica prikladna je za različite analize rečenične strukture - primjerice traženje navodnika.

Lematizacija

Lematizacija teksta obavljena je korištenjem alata CST Lemmatizer[3], točnije modela koji je prilagođen za hrvatski jezik (Agić et al., 2013). Korištenim modelom postiže se preciznost od 98% za lematizaciju hrvatskih riječi.

[2]CroNER - takelab.fer.hr/hr/croner

[3]CST Lemmatizer - cst.dk/online/lemmatiser/uk/

Gramatičko označavanje

Gramatičko označavanje dijelova rečenice ostvareno je korištenjem MST Dependency Parser-a prilagođenog za hrvatski jezik[4] (Agić et al., 2013). Pomoću njega pojavnicama koje čine rečenicu, dobivenim prethodnim rastavljanjem rečenice na pojavnice, pridijeljuje se odgovarajuća gramatička kategorija (subjekt, predikat, itd.).

5.3. Podsustav za ekstrakciju navoda

Podsustav za ektrakciju navoda iz obrađenog teksta implementiran je pomoću metode temeljene na pravilima opisane u poglavlju 3 (Model za ekstrakciju navoda).

U ovom podsustavu prethodno obrađeni tekst se analizira u svrhu pronalaska navoda i njihovih izvora. Zasebno se obrađuje svaka od rečenica iz članka, tako što se uspoređuje s definiranim pravilima, te se u slučaju podudaranja ekstrahiraju navod i govornik. Ako rečenica zadovoljava više pravila, odabire se ono koje je prvo navedeno u popisu pravila iz poglavlja 3. Ključni koraci u analizi rečenice u ovom dijelu su pronalazak navoda (unutar navodnika), pronalazak predikata, provjera pripadnosti predikata skupu izjavnih glagola, te pronalazak subjekta i provjera da li je subjekt imenovani entitet - osoba. Ako bilo koji od navedenih koraka nije zadovoljen ili nije moguće pronaći traženi dio u rečenici, rečenica se odbacuje i obrada se nastavlja s ostalim rečenicama.

Ukoliko je navod zajedno s govornikom uspješno ekstrahiran, obavlja se spremanje podataka u bazu. U bazu se spremaju sljedeći podatci: navod, govornik, link na članak, tekst članka, te teme o kojima članak govori.

5.4. Podsustav za prikaz podataka na grafičkom korisničkom sučelju

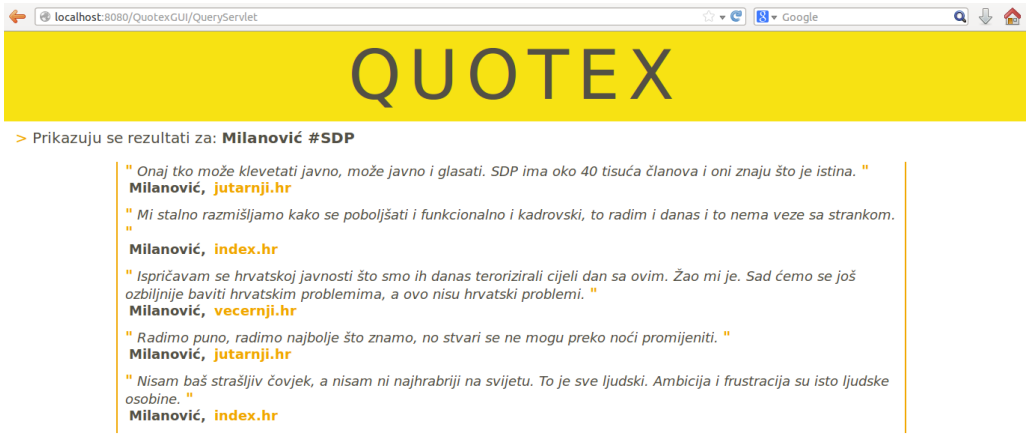
Podsustav za prikaz podataka na grafičkom korisničkom sučelju sastoji se od web-stranice koja pruža korisniku sučelje za jednostavno pretraživanje navoda i baze podataka u kojoj su pohranjeni navodi s pripadajućim entitetima, temama o kojima navodi govore i linkovima na članke iz kojih su navodi ekstrahirani.

5.4.1. Korisničko sučelje

Korisničko sučelje izvedeno je u obliku jednostavne web-stranice putem koje je moguće pretraživati bazu navoda. Navodi se mogu pretraživati po govorniku, ali i po

[4]MST Dependency Parser (prilagođen za hrvatski jezik) - nlp.ffzg.hr/resources/models/tagging/

temama o kojima navodi govore. Uključivanje teme u upit označava se predznačavanjem teme znakom *"#"*, primjerice *"#Sanader"* ukoliko korisnik želi pregledati sve navode koji govore o Sanaderu. Budući da u sustavu nije razriješena anfora između punog imena i prezimena govornika i dijelova istog, pri pretraživanju navoda po govorniku potrebno je posebno pretraživati po punom imenu i prezimenu, i po samom imenu ili samom prezimenu. Također, omogućeno je pretraživanje više tema, pri čemu se kao rezultat prikazuju samo navodi koji govore o objema temama. Kao rezultat upita korisniku se prikazuju traženi navodi i govornik, ali i linkovi na članke iz kojih je navod ekstrahiran. Primjer prikaza rezultata za korisnički upit prikazan je na slici 5.1.



Slika 5.1: Prikaz rezultata za korisnički upit *"Milanović #SDP"*.

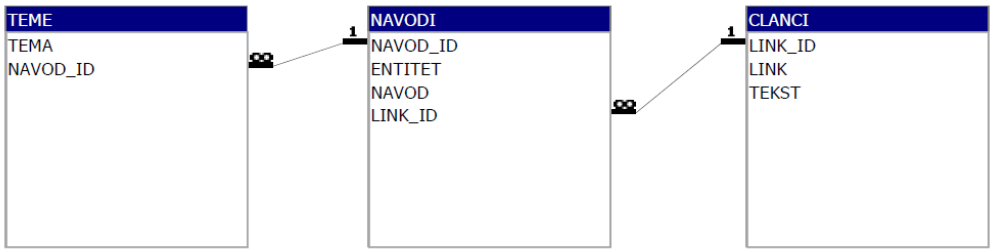
5.4.2. Baza podataka

Baza podataka u kojoj su pohranjeni navodi izvedena je pomoću baze MySQL[5]. Unutar baze navoda izrađene su tri tablice. Prva tablica ("Navodi") sadrži identifikacijske brojeve navoda, nazive entiteta, navode koje su izrekli i identifikacijski broj članka iz kojeg je navod ekstrahiran. Primarni ključ u takvoj tablici je upravo zadnja trojka: entitet-navod-link, budući da su moguće pojave istih entiteta s pripadajućim navodima na različitim internetskim portalima. U drugoj tablici ("Teme") svaka od trojki iz prve tablici povezana je preko identifikacijskog broj navoda s temom o kojoj navod govori, naravno ukoliko su na web-stranici s koje je navod ekstrahiran teme navedene. U trećoj tablici ("Članci") pohranjene su informacije u člancima iz kojih su navodi ekstrahirani, i tako da se u tablici za svaki članak nalazi identifikacijski broj članka, link na članak, te originali tekst članka, koji se koristi u grafičkom korisničkom sučelju u slučaju da

[5]MySQL - www.mysql.com/

link na članak više ne postoji. Ovakvim uređenjem baze podataka omogućeno je pretraživanje po različitim kategorijama: entitetu, entitetu i temi ili samo po temi (ili više njih).

Shema baze podataka prikazana je na slici 5.2. Iz sheme je vidljivo da su teme u tablici tema povezani s navodima preko identifikcijskog broja navoda, dok su navodi s člancima povezani preko identifikacijskog broja članka, koji predstavlja primarni ključ u tablici "Članci".



Slika 5.2: Shema tablica u korištenoj bazi podataka.

6. Zaključak

U završnom radu opisani su različiti pristupi ekstrakciji navoda iz novinskih objava. Proučene su metode ekstrakcije navoda temeljene na pravilima (Pouliquen et al., 2007; De La Clergerie et al., 2011; Krestel et al., 2008), kao i metode strojnog učenja (Elson i McKeown, 2010; O'Keefe et al., 2012). U proučenim radovima, metode temeljene na pravilima uglavnom su korištene za ekstrakciju navoda u obliku upravnog govora, dok su metode strojnog učenja korištene za ekstrakciju navoda u obliku neupravnog govora.

Implementiran je sustav za ekstrakciju navoda iz novinskih objava na hrvatskome jeziku, koji pomoću metode temeljene na pravilima ekstrahira navode u obliku upravnog govora, te ih povezuje s imenovanim entitetima - izvorima navoda. Sustav u ovom trenutku ekstrahira samo navode koji su eksplicitno izrečeni od strane imenovanog entiteta - osobe, dok se slučajevi anafore pri navođenju izvora navoda ili slučajevi izostanka izvora navoda ne ekstrahiraju.

Sustav je vrednovan na temelju ručno označenih navoda i pripadajućih entiteta. Rezultati vrednovanja pokazali su da preciznost tako izgrađenog sustava iznosi 97%, dok je odziv 72%. Neuspješne ekstrakcije u sustavu uzrokovane su najvećim dijelom zbog pogrešaka u predobradi podataka, gdje su dijelovi rečenice pogrešno označeni, pa izgrađena pravila nisu bila zadovoljena.

Izgrađeno je i odgovarajuće grafičko korisničko sučelje, putem kojeg je korisnicima omogućeno pretraživanje zbirke ekstrahiranih navoda prema izvorima navoda i temama o kojima navodi govore. U ovom trenutku, teme prestavljaju oznake na stranici članka koje govore o temi cijelog teksta članka. U budućnosti bi bilo korisno klasificirati svaki navod zasebno u određenu temu, čime bi se povećala preciznost klasifikacije navoda.

Budući da je domena iz koje sustav ekstrahira navode prilično jednostavna i malena, predloženo je proširenje sustava na navode u obliku neupravnog govora, čime bi sustav postao kompleksniji, ali i potpuniji. Za takav oblik ekstrakcije predložene su proučene metode strojnog učenja.

U budućnosti bi bilo dobro razmotriti i druge načine poboljšanja i proširenja sustava. Jedna od značajnijih stavki koja bi bitno uvećala kvalitetu sustava jest razrješava-

nje anafore, koja se pokazala dosta čestim oblikom povezivanja navoda s govornikom. Korištenje alata za rješavanje tog problema znatno bi proširilo domenu unutar koje sustav djeluje.

Ukupno gledano, implementirani sustav postigao je željene rezultate pri ekstrakciji jednostavnih navoda u obliku upravnog govora, ali još uvijek ima prostora za različita poboljšanja koja bi sustav učinila zanimljivijim i potpunijim.

LITERATURA

Željko Agić, Nikola Ljubešić, i Danijela Merkler. Lemmatization and morphosyntactic tagging of croatian and serbian. U *Proceedings of ACL*, 2013.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, i Valentin Tablan. Gate: an architecture for development of robust hlt applications. U *Proceedings of the 40th annual meeting on association for computational linguistics*, stranice 168–175. Association for Computational Linguistics, 2002.

Éric De La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, i Victor Mignot. Extracting and visualizing quotations from news wires. U *Human Language Technology. Challenges for Computer Science and Linguistics*, stranice 522–532. Springer, 2011.

David K Elson i Kathleen McKeown. Automatic attribution of quoted speech in literary narrative. U *AAAI*, 2010.

Goran Glavaš, Mladen Karan, Frane Šaric, Jan Šnajder, Jure Mijic, Artur Šilic, i Bojana Dalbelo Bašic. Croner: A state-of-the-art named entity recognition and classification for croatian language. *Money*, 91:94–15.

Ralf Krestel, Sabine Bergler, René Witte, et al. Minding the source: Automatic tagging of reported speech in newspaper articles. *Reporter*, 1(5):4, 2008.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, i Matthew Honnibal. A sequence labelling approach to quote attribution. U *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, stranice 790–799. Association for Computational Linguistics, 2012.

Bruno Pouliquen, Ralf Steinberger, i Clive Best. Automatic detection of quotations in multilingual news. U *Proceedings of Recent Advances in Natural Language Processing*, stranice 487–492, 2007.

Ekstrakcija navoda iz novinskih objava na hrvatskome jeziku

Sažetak

Završni rad opisuje izgradnju sustava za ekstrakciju navoda iz novinskih objava na hrvatskome jeziku. Model za ekstrakciju navoda i povezivanje navoda s entitetima koji su izrekli navod temeljen je na pravilima. Temelj izgrađenih pravila jest potraga za navodima unutar rečenice, predikatom kojim se povezuje navod s imenovanim entitetom, te subjektom koji je prepoznat kao imenovani entitet. Postignuta preciznost izgrađenog sustava iznosi 97%, dok je odziv 72%. Proučeni su i relevantni radovi iz područja ekstrakcije navoda pomoću metoda temeljenih na pravilima i metoda strojnog učenja. Izgrađeno je i jednostavno grafičko korisničko sučelje kojim se pretražuje baza navoda i pripadnih govornika. U sklopu rada predloženi su i brojni načini poboljšanja i proširenja sustava.

Ključne riječi: Ekstrakcija navoda, povezivanje navoda s entitetom, ekstrakcija informacija pomoću pravila

Quotation Extraction from News Stories in Croatian Language

Abstract

Bachelor's thesis describes an implementation of rule-based, news-wire quotation extraction system for Croatian language. Quotation extraction model is based on rules, each of which tries to find a quote inside a sentence, as well as a predicate which is contained in a set of reported speech verbs and a subject labeled as named entity - person. Precision of implemented quotation extraction system is currently equal to 97%, while recall has value of 72%. Different methods based on rules and machine learning have been studied during the work on the thesis. A simple graphical user interface has been created, for user to search through quotes database. Many ways of system improvement and expansion have been suggested as future work ideas.

Keywords: Quotation extraction, quote attribution, rule-based information extraction