UDC: 004.822:519.76 INFO M: str. 4-9

MODEL SISTEMA ZA EKSTRAKCIJU INFORMACIJA IZ TEKSTOVA PISANIH NA SRPSKOM JEZIKU
MODEL OF A SYSTEM FOR INFORMATION EXTRACTION FROM SERBIAN WRITTEN TEXTS

Staša Vujičić Stanković,
Univerzitet u Beogradu, Matematički fakultet, Studentski trg 16, Beograd

REZIME: U ovom radu je predstavljena narastajuća potreba za nadogradnjom postojećih sistema za ekstrakciju, izdvajanje relevantnih informacija za korisnika, semantičkim podacima i tehnologijama. Takođe su predstavljeni i problemi vezani za obradu sve većeg broja dokumenata u digitalnom obliku pisanih na srpskom jeziku. Pored problema koji potiču iz same prirode srpskog, morfološki bogatog jezika, sagledani su i problemi nedostatka semantičkih resursa i alata za srpski jezik. U skladu sa navedenim, prikazan je razvoj modela sistema za ekstrakciju informacija vođenu ontologijama za tekstove na srpskom jeziku zasnovan na integraciji postojećih resursa koji su razvijeni za obradu kako tekstova na srpskom jeziku, tako i onih koji su namenjeni obradi tekstova na engleskom jeziku; prilagođavanju postojećih tehnika i alata i novim načinima njihove primene u cilju prevazilaženja navedenih problema. Glavni doprinos razvoja ovog modela jeste podsticanje razvoja sistema za izdvajanje informacija relevantnih za korisnika, bazirano na ontologijama, za različite oblasti i primene.

KLJUČNE REČI: ekstrakcija informacija bazirana na ontologijama, semantički veb, obrada prirodnih jezika, srpski jezik.

ABSTRACT: This paper motivates the need for addressing the problem of enriching Information Extraction systems with the Semantic Web data and technologies, stressing the existing problems related to Serbian. Due to the core nature of Serbian, morphologically rich language, which leads to different types of issues in natural language processing tasks and especially Information Extraction process, and the lack of semantic resources and tools for Serbian, it is necessary to develop techniques for overcoming these problems, with particular emphases to take advantage of existing English resources and tools. The model for Ontology-Based Information Extraction is proposed to deal with the imperfection of Information Extraction related to Serbian, and to enhance essential Information Extraction process by incorporating semantic knowledge encapsulated in the ontologies. The scope of this paper is to describe the model based on: integration of existing resources that have been developed for the processing of both texts in the Serbian language, and those that are aimed to be used for processing of texts in English; adapting existing techniques and tools; and inventing new ways of their implementation in order to overcome significant challenges. We believe this model will encourage development of the Ontology-Based Information Extraction systems for specific domains and applications, dealing the increasing volume of data in Serbian.

KEY WORDS: Ontology-Based Information Extraction, Semantic Web, Natural Language Processing, Serbian.

1. UVOD

Ekstrakcija informacija, a posebno ekstrakcija informacija zasnovana na ontologijama, je važan zadatak u razvoju sistema za automatsku obradu sve veće količine informacija u digitalnom obliku, korišćenjem prednosti semantičkog veba. Ovo važi u opštem slučaju, ali predstavlja poseban problem u srpskom jeziku, gde se može prepoznati veliki nedostatak resursa i alata koji će predstavljati temelj za razvoj bilo koje vrste sistema zasnovanog na ontologijama u bilo kojem domenu.

U radu je prikazan model sistema za izdvajanje informacija zasnovan na softverskim sistemima za obradu prirodnih jezika, Unitex[1] i GATE[2], koji u suštini ima za cilj da prevaziđe probleme koji se javljaju pri obradi tekstova na srpskom jeziku. U prvoj fazi se za rešavanje problema morfološke analize primenjuje Unitex sistem i odgovarajući resursi razvijeni za primenu u njemu – elektronski rečnici i kaskade konačnih transduktora. Rezultati se koriste u GATE sistemu za obeležavanje reči adekvatnim kanonskim oblicima, odnosno lemama i Part-Of-Speech (POS) oznakama za vrstu reči. U poslednjoj fazi se koristi WordNet semantička mreža, za mapiranje srpskih i engleskih reči, njihovih lema i sinonima, kako bi se obezbedila mogućnost korišćenja ontologija razvijenih za engleski jezik nad srpskim pisanim tekstovima.

Ovaj rad daje sledeće doprinose: s jedne strane, prikazali smo mogući pristup za prilagođavanje engleskih resursa i alata za ekstrakciju informacija vođenu ontologijama, u cilju poboljšanja postojećih resursa za srpski jezik za upotrebu u složenijim zadacima vezanim za semantički veb; sa druge strane, obezbedili smo temelje za naš budući rad na implementaciji predloženog modela u različitim domenima.

2. EKSTRAKCIJA INFORMACIJA VOĐENA ONTOLOGIJAMA

Ideja o razvoju sistema za ekstrakciju informacija baziranu na ontologijama nije nova [1]. Ipak, postoji očigledna nesrazmera između njenog razvoja kada se posmatra engleski jezik i njenog razvoja koji se odnosi na druge jezike, kao što je srpski. Jedan od naših ciljeva je da iskoristimo resurse razvijene za engleski jezik prilagođavanjem srpskom jeziku.

[1] Unitex sistem: http://www.igm.univ-mlv.fr/~unitex/.

[2] GATE sistem: http://www.gate.ac.uk.

Do sada je urađeno puno na prilagođavanju sistema za ekstrakciju informacija na različite jezike. Učinjeni su izuzetni napori da se razviju resursi i alati za različite zadatke obrade prirodnih jezika vezanih za jezik iznenađenja kao deo DARPA programa *Translingual Information Detection Extraction and Summarization* [2]. Pod jezikom iznenađenja se u ovim eksperimentima podrazumevao nasumično odabran jezik za koji treba razviti resurse i alate slične onima koji postoje za engleski. U eksperimentima sa jezikom iznenađenja za cebuano i hindi jezike, u veoma kratkom roku i bez poznavanja jezika nad kojim se radi, pokazano je da je moguće napraviti višejezične adaptacije i postići veoma kredibilan učinak prilagođenih sistema ([3;4]).

Pored izveštaja o radu i rezultatima na adaptaciji sistema za ekstrakciju informacija za slovenske jezike, kao što su bugarski ili ruski [5], dostupni su i podaci o sličnim eksperimenatima fokusiranim na srpski jezik [6]. Međutim, prilikom ovih adaptacija kao veliki problem istakao se nedostatak odgovarajućih POS tagera za određivanje i označavanje vrsta reči.

Nekoliko autora predstavilo je eksperimente zasnovane na WordNetu i ideji mapiranja engleskog i nekih drugih jezika, kako bi se rešio problem nedostatka dostupnih ontologija na jezicima različitim od engleskog i veliki nedostatak alata za njihovu upotrebu ([7;8]).

Imajući u vidu specifičnosti srpskog jezika, koje će biti opisane u sledećem odeljku, za svrhu razvoja sistema za ekstrakciju informacija vođenu ontologijama, bilo je neophodno prevazići navedene probleme i za svaku reč pre mapiranja utvrditi njenu lemu i vrstu reči.

3. OPIS PROBLEMA KARAKTERISTIČNIH ZA SRPSKI JEZIK

Za rešavanje problema obrade tekstova pisanih prirodnim jezicima kao što su germanski i romanski jezici (a posebno engleski jezik) razvijene su brojne tehnike i alati koji daju dobre rezultate, a takođe su dostupne i velike količine različitih resursa. Za razvoj takvih tehnika, alata i resursa za druge grupe jezika, naročito slovenskih, morfološki bogatih jezika, još uvek ima dosta prostora.

Detaljan pregled karakteristika slovenskih jezika koje dovode do problema pri obradi tekstova pisanih tim jezicima, sa posebnim akcentom na probleme vezane za zadatak izdvajanja informacija, dat je u [9]. Ove vrste problema su evidentne kako u jednostavnijim zadacima ekstrakcije informacija, kao što je prepoznavanje imenovanih entiteta, tako i u složenijim zadacima ekstrakcije informacija, kao što je prepoznavanje odnosa entiteta i prepoznavanje vremenskih odrednica i događaja, to jest, pronalaženje informacija o tome ko je šta kome, kada, gde, kako i zašto.

Problemi koji su posebno izraženi su bogat morfološki sistem, bogata fleksija[3] imenica i slobodan red reči u rečenici. Pored ovih, postoje i problemi složene fleksije vlastitih imena, sa posebnim problemima fleksije vlastitih imena koja potiču iz stranih jezika, kao i problemi vezani za brojeve.

Srpski jezik, kao predstavnik slovenskih jezika, jedan je od morfološki bogatih jezika izuzeno složen za obradu. Ima 7 padeža i 3 vrednosti za brojeve – jedninu, množinu i paukal [10]. Dodatno, lične imenice mogu imati niz izvedenih formi, prideva i priloga. Pored pomenutih problema, među tekstovima pisanim na srpskom jeziku, jednako su zastupljeni oni pisani ćirilicom i latinicom. Ovo predstavlja problem, jer prezapisivanje tekstova u drugo pismo nije jednoznačno u bilo kojoj od standardnih kodnih šema. Primera radi, toponim *New York* u srpskom jeziku pisan latinicom može da se predstavi na dva načina kao *Njujork* (ili čak *NJujork*), dok ćiriličnim pismom može da bude predstavljen kao *Њујорк* ili češće *Њујорк*.

Takođe, pravopis srpskog jezika ne daje preciznu definiciju kako tretirati strana imena koja se neizbežno pojavljuju u tekstovima, kao što su imena ljudi ili organizacija. Na primer, naziv *Microsoft* se pojavljuje u dva oblika – *Mikrosoft* i *Majkrosoft*. Složene reči predstavljaju svojevrstan problem, zbog činjenice da svaka složena reč, ili neki njen deo može da se menja. Na primer, u nazivu *Novi Sad*, oba dela se menjaju: *Novom Sadu*.

Pored ovih problema vezanih za prirodu slovenskih jezika, problem koji je takođe vrlo izražen je nedostatak (slobodnih) resursa koji bi doprineli bržem razvoju tehnika i alata za rešavanje problema različitih vrsta automatske obrade jezika. Koreni ovog problema potiču iz činjenice da, s jedne strane, postoji relativno mali broj govornika pojedinih slovenskih jezika, a da su sa druge strane, ograničena sredstva za ulaganje u razvojnih resursa i alata za jezike koji su manje rasprostranjeni, manje popularni i do sada nedovoljno razvijeni sa stanovišta automatske obrade.

Posebno značajan problem za razvoj sistema za ekstrakciju informacija vođenu ontologijama na srpskom jeziku predstavlja odsustvo ontologija dostupnih na srpskom jeziku i veliki nedostatak srodnih alata i servisa koji bi ih koristili.

4. MODEL SISTEMA ZA EKSTRAKCIJU INFORMACIJA VOĐENU ONTOLOGIJAMA ZA TEKSTOVE PISANE NA SRPSKOM JEZIKU

Iako postoji velika količina resursa, tehnika i alata koji su razvijeni za automatsku obradu tekstova na engleskom jeziku, oni uglavnom ne mogu da se koriste za obradu tekstova napisanih na srpskom jeziku bez prilagođavanja. Postoji sve veća potreba da se omogući njihovo korišćenje za tekstove na srpskom jeziku ili da se razviju slični resursi, tehnike i alati posebno namenjeni obradi srpskog jezika, imajući u vidu ranije navedene probleme.

Ova tvrdnja se odnosi generalno na sve resurse, tehnike i alate koji su namenjeni obradi prirodnog jezika, a u užem smislu, što je od interesa za ovo istraživanje, na njihov razvoj u cilju poboljšanja procesa ekstrakcije informacija iz tekstova na srpskom i njegovo proširenje na ontologije.

[3] Fleksija, oblik, proste ili složene reči je morfosintaksička varijanta reči. Fleksija predstavlja obrazovanje oblika reči kombinacijom (gramatičke) osnove i vezanih gramatičkih morfema (nastavaka za oblik i infiksa) [10].

Model koji će biti predstavljen je razvijen sa ciljem da se prevaziđu navedeni problemi. Zasniva na dva sistema za obradu prirodnog jezika, sistemima Unitex i GATE, i različitim resursima, koji su razvijeni od strane Grupe za obradu prirodnih jezika na Matematičkom fakultetu Univerziteta u Beogradu (elektronskim rečnicima, korpusima, konačnim transduktorima, semantičkoj mreži WordNet za srpski jezik...) i adaptirani za potrebe ovog istraživanja.

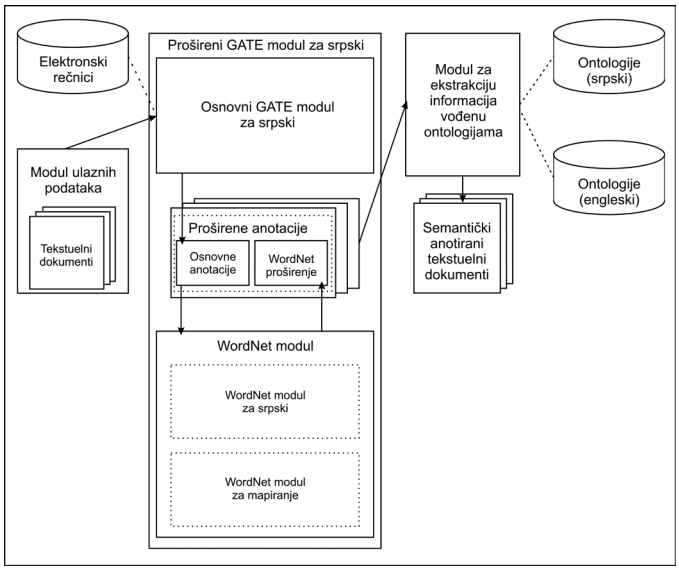
Arhitektura modela sistema za ekstrakciju informacija vođenu ontologijama za tekstove na srpskom jeziku prikazana je na Slici 1. Ulazne podatke sistema čine elektronski rečnici i korpus tekstova na srpskom jeziku. Centralni deo sistema uključuje prošireni modul za obradu tekstova na srpskom jeziku i modul za ekstrakciju informacija vođenu ontologijama koji proizvodi izlazne podatake – dokumente sa odgovarajućim semantičkim oznakama.

4.1. *Modul ulaznih podataka*

Modul ulaznih podataka se sastoji od dokumenata u formatu čistog teksta. Ovi dokumenti su prikupljeni iz više različitih izvora i pisani su na srpskom jeziku. Korpusi generalno sadrže tekstove novinskih članaka, dela iz klasične i savremene književnosti, tekstova preuzetih sa interneta – na primer članaka preuzetih sa Vikipedije, mikro-tekstuelnih unosa sa Tvitera i tako dalje.

4.2. *Prošireni GATE modul za srpski jezik*

Prošireni GATE modul čine dva dela: osnovni GATE modul za srpski jezik i WordNet modul.



Slika 1. – *Arhitektura modela sistema za ekstrakciju informacija vođenu ontologijama za tekstove pisane na srpskom jeziku.*

4.2.1. Osnovni GATE modul za srpski jezik

Osnovni GATE modul za srpski jezik razvijen je kao modul (GATE komponenta) za obradu tekstova na srpskom jeziku u sistemu GATE. Modul je zasnovan na funkcionalnostima razvijenim u sistemu Unitex – elektronskim rečnicima i konačnim transduktorima, da bi se omogućilo izvršavanje osnovnih zadataka (pre)procesiranja tekstova kao što su segmentiranje rečenica, tokenizacija, obeležavanje vrsta reči i morfološka analiza (lematizacija).

Unitex sistem

Unitex je besplatni softverski sistem otvorenog koda za obradu korpusa, zasnovan na automatima, namenjen za primenu elektronskih rečnika i gramatika na tekstove. Jedna od glavnih karakteristika Unitex sistema je da omogućava obradu tekstova na više jezika, uključujući i srpski, pružajući mogućnost postavljanja znatno kompleksnijih upita nego što su jednostavni regularni izrazi nad karakterima.

Unitex sistem koristi elektronske rečnike DELA tipa (*Dictionnaires Electroniques du LADL* ili *LADL electronic dictionaries*). DELA format za morfološke rečnike sadrži informacije koje omogućavaju rešavanje problema segmentacije, kao i morfološke, sintaksičke i semantičke obrade teksta.

Elektronski rečnici u DELA formatu su organizovani u sistem. Sistem DELA morfoloških rečnika se sastoji od rečnika: DELAS (*Dictionnaires électroniques des mots simples*) – rečnik lema prostih reči (kanonskih oblika reči), DELAF (*Dictionnaires électroniques des formes fléchies*) – rečnih svih oblika reči svih lema iz rečnika DELAS, kao i sličnih rečnika lema i oblika složenih reči: DELAC (*Dictionnaire électronique des mots composés*) i DELACF (*Dictionnaire électronique des mots composés fléchis*). Više o DELA formatu može se naći u [12].

Srpski morfološki rečnik prostih reči, DELAS, sadrži 130 000 lema, od čega se većina odnosi na generalnu leksiku, dok preostale leme predstavljaju različite vrste ličnih imena [13;14]. Sa druge strane, DELAF rečnici sadrže približno 1 450 000 različitih oblika reči. Veličine DELAC i DELACF rečnika su redom približno 10 500 i 54 000 lema [15].

Opisani morfološki rečnici, zajedno sa nizom konačnih transduktora koji su razvijeni za obradu srpskog jezika u sistemu Unitex, predstavljaju dobru osnovu za obradu tekstova pisanih na srpskom jeziku. Ipak, u sistemu Unitex, ne postoji podrška za rad sa ontologijama. Jedan od glavnih ciljeva u ovom istraživanju bio je da se iskoristi najbolje iz Unitexa u sistemu GATE, kroz rešavanje problema morfološke analize i prepoznavanje složenih reči, kreiranjem odgovarajućih elektronskih rečnika i pravilnim anotiranjem tekstova.

GATE sistem

GATE (*General Architecture for Text Engineering*) je besplatan softver otvorenog koda. U osnovi , to je arhitektura, okvir i razvojno okruženje za obradu prirodnih jezika. Jedna od ključnih karakteristika sa stanovišta ovog istraživanja je mogućnost sistema GATE da se nadograđuje kako bi se podržala obrada tekstova pisanih na bilo kom prirodnom jeziku.

Resursi sistema GATE koji su važni za zadatak ekstrakcije informacija, integrisani su u sistem ANNIE (*A Nearly-New Information Extraction*). Detaljan prikaz sistema ANNIE može se pogledati u [16]. Glavna prednost organizacije GATE arhitekture je da se svaki resurs može individualno koristiti ili kombinovati sa novim resursima koji su uključeni u sistem, a

namenjeni su radu sa određenim jezicima ili specifičnim domenima. Sledeći ANNIE resursi su od posebnog značaja za ovo istraživanje:

- *ANNIE tokeniser* koji deli tekst na tokene, i prepoznaje reči, brojeve, simbole, beline i znakove interpunkcije. Ovaj resurs se u najvećem broju slučajeva može koristiti za različite jezike bez (značajnih) izmena. Ovo tvrđenje takođe važi i za srpski jezik.
- *ANNIE sentence splitter* koji je zasnovan na kaskadi konačnih transduktora i deli tekst na rečenice. Najčešće ne zahteva velike izmene da bi se koristio za različite jezike ili za različite tipove aplikacija, što takođe važi i za srpski.
- *ANNIE POS tagger* koji svakoj reči pridružuje informaciju o vrsti te reči (POS tag) u formi anotacije. Ovaj resurs je zavisan od jezika na koji se primenjuje i jedan od glavnih ciljeva ovog istraživanja je upravo njegovo prilagođavanje srpskom jeziku.

GATE modul za srpski jezik

Primena resursa koji su originalno uključeni u GATE sistem ne daje podjednako dobre rezultate pri obradi tekstova pisanih na srpskom jeziku u individualnim fazama: dok faze u kojima se vrši identifikacija tokena i segmentacija daju zadovoljavajuće rezultate, zbog čega ih je moguće izvršiti bez modifikacija, isto se ne može reći i za pridruživanje odgovarajuće leme i određivanje vrste reči, gde je neophodno izvršiti modifikacije.

Osnovna verzija POS tagera u GATE sistemu zasnovana je na Brilovom tageru, koji kao i drugi statistički POS tageri ne daje rezultate zadovoljavajuće tačnosti prilikom prilagođavanja i primene nad srpskim jezikom [17].

Imajući na umu navedene nedostatke, naša ideja za rešavanje navedenih problema je da se proizvedu odgovarajući DELA rečnici za tekstove koji se obrađuju, upotrebom sistema Unitex i njegovih funkcionalnosti i resursa, a potom da se izvrši njihovo prilagođavanje za upotrebu u sistemu GATE. U skladu sa tim, nad tekstovima se pored osnovne faze preprocesiranja dokumenata upotrebom *ANNIE sentence splitter* i *ANNIE tokeniser* komponenata vrši preprocesiranje upotrebom ugrađenih funkcionalnosti Unitex sistema. Time se obezbeđuje formiranje odgovarajućih DELA elektronskih rečnika i povezivanje svake reči dokumenta sa odgovarajućom lemom i vrstom reči, koje su vrednosti standardnih GATE anotacija *lemma* i *POS category* neophodnih za rad ostalih modula sistema.

Preciznije, DELA rečnici su predstavljeni u formatu čistog teksta. Svaka linija u rečiku predstavlja podatke o reči iz ulaznog teksta – lemi (kanonskom obliku) te reči i dodatnim gramatičkim, semantičkim i flektivnim informacijama. Jedan primer unosa, jedna linija iz rečnika prostih reči, je *kući,kuća.N:fs3q:fs7q*. Prvi deo (*kući*) je obavezan i predstavlja oblik proste reči pronađen u ulaznom tekstu. Drugi deo (*kuća*) predstavlja lemu (kanonski oblik) ulazne reči. Treći deo (*N:fs3q:fs7q*) predstavlja gramatičke informacije i informacije o kodu koji označava klasu lema sa istim flektivnim svojstvima – *N* označava imenicu, *f* označava da je ova imenica ženskog roda, a *s* označava da je u pitanju imenica u jednini. Podaci iz ovog formata odgovaraju vrednostima GATE anotacija *lemma* i *POS category* (u ovom primeru , *lemma = kuća*, *POS category = N*).

Dakle, nakon primene GATE modula za srpski jezik, ulazni tekst će biti dopunjen osnovnim skupom anotacija, to jest, svaka reč iz teksta će biti povezana sa odgovarajućom anotacijom koja sadrži obeležja – leme i vrste reči (*lemma* i *POS category*). Takav skup je potrebno obezbediti kao ulazne podatke za WordNet modul.

4.2.2. WordNet modul

Ranije je istaknut nedostatak ontologija koje su razvijene za srpski jezik kao jedan od glavnih problema u savremenoj obradi tekstova pisanih na srpskom jeziku. Da bi se ovaj problem prevazišao, neophodno je da se obezbedi neka vrsta srpsko-engleskog mapiranja koje bi omogućilo upotrebu ontologija razvijenih za engleski. Predloženo rešenje se zasniva na upotrebi WordNet višejezične semantičke mreže.

Semantička mreža WordNet je za srpski jezik inicijalno bila razvijena u okviru projekta BalkaNet (*A multilingual semantic network for the Balkan Languages*), koji je imao za cilj kreiranje semantičkih i leksičkih mreža balkanskih jezika [18] i njihovu integraciju u globalni WordNet [19;20].

WordNet modul vrši dopunu osnovnih anotacija odgovarajućim sinonimima i mapiranjem (prevođenjem) leme na srpskom jeziku i lema na engleskom jeziku, konsultujući višejezičnu semantičku mrežu. Ove anotacije, zajedno sa prethodno dodeljenim anotacijama sa obeležjima lema i vrsta reči, su od suštinskog značaja za dalju obradu u narednom modulu.

4.3. *Modul za ekstrakciju informacija vođenu ontologijama*

Ontologije kao opisi koncepata i odnosa koji mogu da postoje u njihovoj hijerarhiji, su od ključnog značaja za razvoj semantičkog veba, kao i za unapređenje tradicionalnih tehnika ekstrakcije informacija i alata. Ovo važi u opštem slučaju, ali predstavlja poseban problem u srpskom jeziku, gde se prepoznaje nedostatak takvih resursa koji bi bili osnova za dalji razvoj bilo koje vrste sistema zasnovanih na ontologijama u bilo kom domenu.

Da bi se prevazišao ovaj problem, modul za ekstrakciju informacija vođenu ontologijama, zasnovan je na funkcionalnostima sistema GATE vezanim za ontologije. Modul će u svojoj implementaciji obavljati semantičko označavanje tekstova, korišćenjem domenski specifičnih ontologija razvijenih za engleski jezik, uzimajući u obzir da su sve reči iz srpskih dokumenata (i njihovi sinonimi) mapirani u odgovarajuće engleske reči u okviru proširenog GATE modula za srpski. Tu bi u budućnosti postojala mogućnost da se izvrši semantička anotacija upotrebom ontologija za srpski ili bilo koji drugi jezik.

5.JEDNA PRIMENA PREDLOŽENOG MODELA SISTEMA ZA EKSTRAKCIJU INFORMACIJA VOĐENU ONTOLOGIJAMA

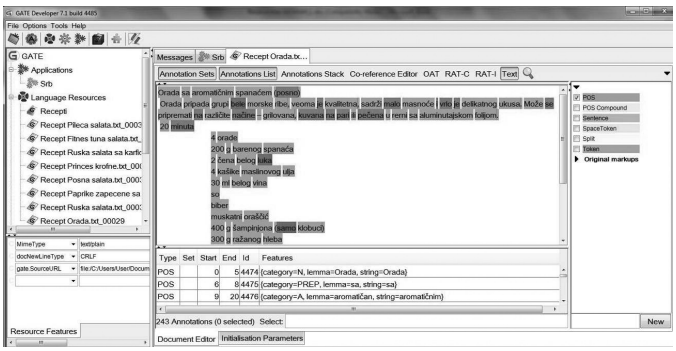
Jedan od projekata u okviru koga je bilo neophodno prilagođavanje postojećih alata kako bi se uzele u obzir specifičnosti srpskog jezika jeste izdvajanje specifičnih informacija iz

kulinarskih recepata. U tu svrhu iskorišćen je i implementiran model opisan u ovom radu. Cilj je da se iz korpusa recepata omogući izdvajanje relevantnih koncepata, osobina i relacija, kako bi se pored standardnog pretraživanja korpusa recepata po ključnim rečima, omogućilo naprednije pretraživanje po složenijim kriterijumima i postavljanjem složenijih upita od strane korisnika (na primer, pretraživanje korpusa recepata prema broju kalorija, iako ta informacija nije eksplicitno navedena u samim receptima, ili pretraživanje po kriterijumima vezanim za navođenje određenih alergija na hranu ili bolesti zbog kojih korisnik ne sme da konzumira određene vrste namirnica pa je neophodno uključiti i supstitucije namirnica, itd.).

Prilikom implementacije sistema za konkretan domen primene neophodno je prilagoditi svaki od opisanih modula. Tokom postupka prilagođavanja i korišćenja modela u kulinarskom domenu, prepoznati su dodatni problemi i pravci razvoja koji će biti rešavani u našem daljem radu.

Modul ulaznih podataka u kulinarskom domenu sadrži tekstove recepata koji se prikupljaju u formi čistih tekstuelnih dokumenata, a skladište u internoj reprezentaciji korpusa sistema GATE. S obzirom na narastajuću količinu kulinarskih sadržaja, kako recepata, tako i različitih saveta i opisa, korpus je formiran preuzimanjem tekstova sa veba.

Za preuzimanje tekstova sa veb strana postoje brojni besplatni programi poput BootCaT[4]-a, koji daju zadovoljavajuće rezultate. Da bi se na što bolji način iskoristila izvorna struktura veb strana, a ne samo tekst koji se korisnicima prikazuje, prilikom implementacije ovog modela razvijeni su programi prilagođeni stranama sa kojih je preuziman sadržaj i meta-podaci koji se mogu iskoristiti u kasnijoj fazi izgradnje ontologije. Programi su napisani u programskom jeziku Java koji pruža podršku za obradu tekstova upotrebom regularnih izraza. Na ovaj način kreiran je korpus od približno 14000 recepata pisanih latinicom, preuzetih sa vodećih domaćih sajtova kulinarskog domena kao što su Recepti[5], Kuhinjica[6], Veliki kuvar[7] i slično. S obzirom da je za potrebe narednog modula potrebno da tekstovi budu u čistom tekstuelnom formatu, naknadno su uklonjeni meta-podaci, kao i recepti koji su pisani bez upotrebe dijakritičkih znakova (gde je „š" pisano kao „s", „č" i „ć" kao „c", „ž" kao „z", a „đ" kao „d").



Slika 2. – *Rezultat rada osnovnog GATE modula za srpski jezik. Pregled skupa anotacija i njihov oblik u slučaju prostih reči.*

4 BootCaT: http://bootcat.sslmit.unibo.it

5 Recepti: http://www.recepti.com

6 Kuhinjica: http://www.kuhinjica.rs

7 Veliki kuvar: http://velikikuvar.com

U semantičkoj mreži WordNet reči kojima se opisuje isti pojam organizovane su u koncepte, odnosno skupove sinonima – sinsete (eng. *synset*). Svaki sinset je jednoznačno određen jedinstvenim identifikacionim brojem, a za svaku reč u sinsetu je između ostalog navedena i njena lema – literal. Prilikom implementacije **WordNet modula** za mapiranje lema na srpskom jeziku u odgovarajuće leme i sinonime na engleskom jeziku razvijen je poseban program pisan u programskom jeziku Java. Program je zasnovan na činjenici da su u svim WordNet semantičkim mrežama razvijenim za različite jezike, isti koncepti označeni istim jedinstvenim identifikacionim brojem. Na taj način se za lemu svake reči teksta, dobijenu radom osnovnog GATE modula za srpski jezik, konsultuje WordNet za srpski jezik, utvrđuje se jedinstveni identifikacioni broj sinseta kome ta lema pripada, vrši preuzimanje svih literala iz odgovarajućeg sinseta engleskog WordNeta i potom se oni pridružuju početnoj reči u vidu anotacija.

Prilikom razvoja i testiranja ovog dela sistema, pokazalo se da postoji nekoliko problema, od kojih se najviše ističe nedovoljna zastupljenost kulinarskih pojmova (naziva namirnica, pribora itd.), kako u srpskom, tako i u engleskom WordNetu. Jedan od glavnih zadataka autora tokom daljeg razvoja sistema biće domensko dopunjavanje WordNeta.

Kako je ranije pomenuto, za rad **modula za ekstrakciju informacija vođenu ontologijama**, u sistemu GATE je implementiran niz funkcionalnosti, koje tokom rada koriste opisane anotacije i domenske ontologije. Iako je na opisani način problem nedostatka domenskih ontologija razvijenih za srpski jezik moguće rešiti upotrebom domenskih ontologija razvijenih za engleski jezik, pokazalo se da u kulinarskom domenu ne postoji kompletna, javno dostupna ontologija. Koliko je autorima poznato, u literaturi postoje opisi domenskih ontologija koje predstavljaju ekspertsko znanje iz ove oblasti, ali nijedna od njih ne može da se upotrebi u okviru ovog sistema. Neke od ontologija su previše specifične – kakve su ontologije vina predstavljene u [21] i [22], druge su previše pojednostavljene ili nisu otvorenog koda – kakva je ontologija kulinarskih recepata predstavljena u [23] ili PIPS (*Personalized Information Platform for Health and Life Sciences*) ontologija [24], tako da se za upotpunjavanje implementacije sistema radi na razvoju odgovarajuće domenske ontologije.

6. ZAKLJUČAK

U ovom radu smo predstavili prednosti koje nastaju integracijom oblasti ekstrakcije informacija i oblasti semantičkog veba, opšte probleme koji nastaju pri realizaciji te ideje i posebno probleme vezane za srpski jezik. Predložili smo model sistema za izdvajanje informacija zasnovano na ontologijama iz tekstova pisanih na srpskom jeziku, koji se zasniva na upotrebi i prilagođavanju resursa i alata razvijenih za engleski jezik.

Sistemi za ekstrakciju informacija su uvek vezani za određeni domen, pa bi kvalitet njihovog rada morao da se ilustruje i proceni kroz konkretne primene u specifičnim obla-

stima. Stoga, predloženi model sistema ne može da se oceni u opštem slučaju, jer evaluacija njegovih performansi zavisi od specifičnih resursa, kao što su korpusi ili ontologije.

Procena će biti data u našem daljem radu kroz primenu modela na konkretan domen i ocenu rezultata te implementacije. U tom procesu će se koristiti klasične mere performansi za procenu (kao što su odziv, preciznost i F-mera), ali je neophodno da se koriste i mere uvedene konkretno za evaluaciju sistema za ekstrakciju informacija vođenu ontologijama (kao što su dodatna preciznost – *Augmented Precision* i dodatni odziv – *Augmented Recall* [25]).

Očekivani doprinosi ovog rada su specifične implementacije predloženog modela sistema za različite domene. Ove implementacije će poboljšati ekstrakciju informacija upotrebom semantičkog znanja sadržanog u razvijenim ontologijama za engleski jezik za različite oblasti. Time će se obezbediti da se sve veća količina sadržaja pisanih na srpskom jeziku koja se može pronaći na vebu, približi novoj generaciji veba – semantičkom vebu.

LITERATURA

[1] Maynard, D., H. Saggion, M. Yankova, K. Bontcheva, & W. Peters (2007). Natural language technology for information integration in business intelligence. In *10th International Conference on Business Information Systems*, April 25-27, 2007, Poland, ed. W. Abramowicz, editor, URL: http://gate.ac.uk/sale/bis07/musing-bis07-final.pdf.|

[2] Oard, D. (2003). The surprise language exercises. In *ACM Transactions on Asian Language Information Processing (TALIP)*, v.2 n.2, pp.79-84, June 2003.

[3] Yarowsky, D. (2003). *Scalable elicitation of training data for machine translation*. Team TIDES. URL: http://language.cnri.reston.va.us/TeamTIDES.html.

[4] Maynard, D., V. Tablan & H. Cunningham (2003). NE recognition without training data on a language you don't speak. In *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, Sapporo, Japan.

[5] Maynard, D. & H. Cunningham (2003). Multilingual adaptations of a reusable Information Extraction tool. In *Proceedings of the 10th Conference of the European Chapter of the Association of Computational Linguistics* (*EACL 2003*).

[6] Vujičić Stanković, S. (2012). Named Entity Recognition in the System for Information Extraction. In *Selected Papers from SinFonIJA 3*, eds. Sabina Halupka-Rešetar, Maja Marković, Tanja Milićev i Nataša Milićević, pp. 206-223, Cambridge Scholars Publishing.

[7] Krstev, C., R. Stankovic, Vitas, D. & I. Obradovic (2008). The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines. In *6th LREC International Conference on Language Resources and Evaluation*, Marrakech, Marocco.

[8] Fellbaum, C. & P. Vossen (2012). *Challenges for a multilingual wordnet*. Springer, Netherlands. URL: http://dx.doi.org/10.1007/s10579-012-9186-z.

[9] Przepiórkowski, A. (2007). Slavic Information Extraction and Partial Parsing. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pp.1–10, Association for Computational Linguistics, Prague, Czech Republic, URL: http://www.aclWeb.org/anthology/W/W07/W07-1701.

[10] Stanojčić, Z. & Lj. Popović (2002). *Gramatika srpskoga jezika*. Zavod za udžbenike i nastavna sredstva, Beograd.

[11] Krstev, C. & D. Vitas (2007). The Treatment of Numerals in Text Processing. In *Proceedings of 3nd Language & Technology Conference*, October 5-7, 2007, Pozna, Poland, ed. Zygmunt Vetulani, pp. 418-422, IMPRESJA Widawnictwa Elektroniczne S.A., Pozna.

[12] Courtois, B. & M. Silberztein (1990). *Dictionnaires électroniques du français*. Langue française 87, Paris, Larousse.

[13] Krstev, C., D., Vitas, I., Obradović & M. Utvić (2011). E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, *FSMNLP* 2011, pp. 48-56.

[14] Gucul-Milojević, S. (2010). *Personal Names in Information Extraction*. INFOtheca 11, 1 (April 2010), 53a-63a.

[15] Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade, Belgrade.

[16] Cunningham, H., D. Maynard, K. Bontcheva & V. Tablan (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (*ACL'02*).

[17] Popović, Z. (2008). *Evaluation of the POS tagging programs for the annotation of the Serbian texts*. Faculty of Mathematics, University of Belgrade, Belgrade.

[18] Stamou, S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit & M. Grigoriadou (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the International Wordnet Conference*, pp. 12-14, 21-25 January 2002, Mysore, India.

[19] Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

[20] Miller G. A. (1990). Introduction to WordNet: An On-Line Lexical Database. In *International Journal of Lexicography*, Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. Vol. 3, No. 4, 1990, 235-244.

[21] Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. Knowledge Systems Lab./Stanford Medical Informatics, Stanford Univ. Technical Report. KSL-01-05/SMI-2001-0880.

[22] Graça, J., Mourão, M., Anunciação, O., Monteiro, P., Pinto, H. S., & Loureiro, V. (2005). Ontology building process: the wine domain. In *Proceedings of the 5th Conference of EFITA*.

[23] Villarías, L. G. (2004). *Ontology-based semantic querying of the web with respect to food recipes*. Informatics and Mathematical Modelling, Technical Univ. of Denmark. Lyngby, Denmark. Master Thesis. ISSN 1601-233X.

[24] Cantais, J., Dominguez, D., Gigante, V., Laera, L., & Tamma, V. (2005). An example of food ontology for diabetes control. In *Proceedings of the International Semantic Web Conference 2005 Workshop on Ontology Patterns for the Semantic Web*, Galway, Ireland.

[25] Maynard, D., W. Peters & Y. Li (2006). Metrics for evaluation of Ontology-Based Information Extraction. In: *Proceedings of the WWW 2006 Workshop on Evaluation of Ontologies for the Web*, ACM, New York.

M. Sc. Staša Vujičić Stanković, Univerzitet u Beogradu, Matematički fakultet, Studentski trg 16, Beograd
mail: stasa@matf.bg.ac.rs
Oblasti interesovanja: obrada prirodnih jezika, ekstrakcija informacija, semantički veb, teorija formalnih jezika i automata