Proširivanje upita zasnovano na leksičkim resursima

Ranka Stanković, Ivan Obradović, Cvetana Krstev

PROŠIRIVANJE UPITA ZASNOVANO NA LEKSIČKIM RESURSIMA

QUERY EXPANSION SUPPORTED BY LEXICAL RESORCES

Ranka Stanković, *Univerzitet u Beogradu, Rudarsko-geološki fakultet*
Ivan Obradović, *Univerzitet u Beogradu, Rudarsko-geološki fakultet*
Cvetana Krstev, *Univerzitet u Beogradu, Filološki fakultet*

Sadržaj - *U radu je opisano kako se leksički resursi za srpski jezik i softverski alati, razvijeni u okviru Grupe za jezičke tehnologije Univerziteta u Beogradu, mogu koristiti za unapređenje postavljanja upita. Rezultati pretrage mogu biti značajno unapređeni korišćenjem različitih leksičkih resursa, kakvi su morfološki rečnici i semantičke mreže. Izloženi pristup može se iskoristiti i u Sistemu naučnih, tehnoloških i poslovnih informacija, jer je efikasno pretraživanje ovog dragocenog resursa, imajući u vidu njegovu heterogenost i obim, kao i preovladavajući tekstualni sadržaj, veoma složen zadatak. U radu se predstavlja i softverski alat WS4LR koji je razvijen i koristi se za rešavanje raznih zadataka u Grupi, i web aplikacija WS4QE koja, zajedno sa pratećim veb servisima, omogućava da se niz zadataka sada reši i preko veba. Pored kratkog opisa nekih od jezičkih resursa za srpski jezik, biće opisano kako se funkcije alata WS4LR mogu koristiti za njihovo održavanje i razvoj, kao i neke od mogućnosti za proširenje upita na web-u i korišćenje tih proširenih upita, koje pruža web aplikacija WS4QE.*

Abstract - *This paper presents how resources and tools developed within the Human Language Technology Group at the University of Belgrade can be used for improvement of queries. Search results can be substantially improved by using various lexical resources, such as morphological dictionaries and semantic networks. The outlined approach may be used within the System of scientific, technical and business information. Efficient exploration of such a valuable source, in view of the diversity and amount of archived data, including textual, is a complex task. The tools we describe are WS4LR, a software tool that has already been developed and used for solving different tasks within the Group, and a web application named WS4QE, accompanied by several web services, that enables the solution of various tasks via the web. Besides a short description of the lexical resources for Serbian involved, we shall also describe how the functions of the WS4LR tool can be used for their maintenance and development, as well as some possibilities for web query expansion offered by the WS4QE web application and the use of these expanded queries.*

1. UVOD

Upit koji se postavlja ka velikim tekstuelnim kolekcijama se u aktuelnim sistemima za pronalaženje informacija svodi na različite varijante poređenja karakterskih niski. Pri tome se ni ključne reči, ni sami tekstovi ne tretiraju kao objekti koji su organizovani prirodnim jezikom. Proširivanjem upita zasnovanom na leksičkim resursima mogu se dobiti znatno bolji rezultati prilikom pretraživanja ovih tekstualnih kolekcija. Tako se i pretraživanje u Sistemu naučnih, tehnoloških i poslovnih informacija, imajući u vidu njegovu heterogenost i obim, kao i preovladavajući tekstualni sadržaj može značajno unaprediti korišćenjem različitih leksičkih resursa.

U jezicima sa bogatim morfološkim sistemom, kakav je srpski jezik, standardni način postavljanja upita ispoljava dodatne manjkavosti. Sem toga, jedan isti koncept može biti leksikalizovan na više načina, odnosno pomoću više sinonima koji ga označavaju, o čemu prosečan korisnik prlikom postavljanja standardnog upita obično ne vodi računa. Upit se takvim situacijama može proširiti korišćenjem sinonimije, ali isto tako i na osnovu drugih semantičkih relacija (npr. koristeći relacije podređenosti i nadređenosti). Konačno, koncepti se leksikalizuju na različite načine u različitim jezicima a relevantan dokument za postavljeni upit može biti i dokument na jeziku koji se razlikuje od jezika upita, pa stoga i višejezično proširenje upita može biti od značaja.

2. LEKSIČKI RESURSI

Leksički resursi za srpski jezik se razvijaju u okviru Grupe za jezičke tehnologije na Matematičkom fakultetu Univerziteta u Beogradu već duži niz godina, tako da je danas na raspolaganju veliki broj različitih resursa, razvijenih u značajnom obimu (Vitas et al., 2003). Pored korpusa srpskog jezika, kao i višejezičnih paralelnih korpusa, od posebnog su značaja sistem morfoloških rečnika srpskog jezika, kao i semantička mreža za srpski jezik (srpski wordnet) razvijena u okviru međunarodnog projekta Balkanet (Tufiş ed., 2004).

Srpski elektronski morfološki rečnik sadrži preko 117,000 lema prostih reči i 1,400,000 odgovarajućih flektivnih oblika

i razvijen je u LADL formatu.[1] Morfološki rečnici u istom formatu postoje za mnoge druge jezike, uključujući francuski, engleski, grčki, portugalski, ruski, tai, korejanski, italijanski, španski, norveški, arapski, nemački, poljski i bugarski. Elektronski morfološki rečnik složenih reči je u intenzivnom razvoju i trenutno ima preko 3000 odrednica.

Wordnet, semantička mreža reči, predstavlja leksičku bazu podataka kojom se realizuje semantička mreža koncepata za određeni jezik, a zasniva se na predstavljanju svakog koncepta pomoću skupa sinonimnih parova reč-značenje. Od ovh parova gradi se osnovni element ove baze - sinset (*synset*, od engleskog *synonymous set*). Upotreba para reč-značenje se zasniva na pristupu koji se koristi u klasičnim rečnicima govornog jezika, gde jednoj reči odgovara više mogućih značenja, koja se na poseban način obeležavaju. U samoj bazi podataka *wordnet* svaki sinset pored samih parova reči-značenje sadrži i druge podatke, od kojih su najznačajniji oznaka vrste reči - POS, zatim definicija koncepta, primeri upotrebe reči iz sinseta za označavanje tog koncepta i semantičke relacije koje ga povezuju sa drugim sinsetima.

Prva mreža reči za engleski jezik pod nazivom Prinstonski wordnet (PWN) razvijena je 1985. godine u *Cognitive Science Laboratory* Prinstonskog univerziteta (Fellbaum, 1998). *EuroWordNet (EWN)* je bio projekat u okviru koga su pored mreže reči za engleski razvijene odovarajuće mreže za još sedam evropskih jezika: holandski, italijanski, španski, francuski, nemački, češki i estonski. Sve mreže u okviru EWN razvijane su po ugledu na PWN ali je EWN uveo i jednu značajnu novinu: između sinseta kojima je isti koncept predstavljen u različitim jezicima u EWN je uspostavljena veza preko tzv. međujezičkog indeksa (Inter-Lingual-Index ili skraćeno ILI).

Polazeći od istog principa, u okviru projekta *BalkaNet*, koji je od 2001. godine do 2004. godine finansirala Evropska komisija, razvijene su mreže reči za bugarski, grčki, rumunski, turski i srpski jezik, a nastavljen je razvoj mreže reči za češki, koji je započet u okviru *EuroWordNet* projekta (Tufiş, 2004). U projekat *BalkaNet* je bilo uključeno 13 istraživačkih i naučnih institucija iz zemalja za čije su jezike mreže razvijane, ali i iz Francuske i Holandije. Za svaki jezik formiran je nacionalni razvojni tim, koji je u slučaju srpskog jezika predstavljala Grupa za jezičke tehnologije Univerziteta u Beogradu. Po završetku ovog projekta, razvoj SWN je nastavljen i ova mreža reči danas sadrži blizu 25000 parova reč-značenje organizovanih raspoređenih u nešto manje od 15000 sinseta.

Pored već pomenutih resursa, u Grupi se koriste i razvijaju i grafovi, koji se u lingvističkim softverima koriste za formalizaciju lingvističkih fenomena i za obradu (parsiranje) teksta. Pored njih, i dvojezične, paralelne liste, kao pomoćni resurs pri pretraživanju i prevođenju. Konačno, Grupa učestvuje i u razvoju višejezične ontologije vlastitih imena (Prolex) (Maurel, 2007), organizovane oko koncepta vlastitog imena, kao jedinstvenog koncepta u različitim jezicima. Naime, u višejezičnom kontekstu, opis vlastitih imena se ne može svesti samo na elektronski rečnik, zbog kompleksnosti semantičkih veza kojih ih povezuju.

[1] LADL format, naziv potiče od naziva laboratorije u kojoj je ovaj pristup obradi prirodnih jezika nastao: Laboratoire d'Automatique Documentaire et Linguistique

3. WS4LR

Sa rastom broja, obima i sadržaja resursa, pojavila se potreba za razvojem jednog softverskog alata kojim bi se olakšalo njihovo održavanje, korišćenje i integracija i omogućio dalji efikasan razvoj. Pored različitih formata resursa, poseban problem bili su i različiti kodni rasporedi koji su se vremenom javljali u resursima. Da bi se rešili ovi problemi heterogenosti nastalo je integrisano i prilagodljivo softversko rešenje, nazvano WS4LR (Work Station for Lexical Resources - Radna stanica za leksičke resurse) kojim je omogućeno upravljanje i rad sa pojedinačnim resursima, kao i njihovo integrisanje (Krstev et al., 2006).

Slika 1. Rukovanje resursima u WS4LR

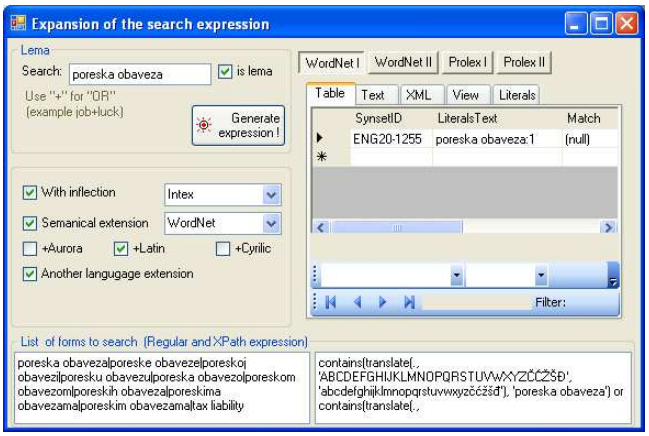
WS4LR se sastoji od više podsistema, od koji su najznačajniji:

- podsistem za održavanje sistema morfoloških elektronskih rečnika prostih i složenih reči,
- podsistem za razvoj i unapređenje wordnet-a, koji podržava kako rad sa pojedinačnim wordnet-ovima tako i sinhronizovano korišćenje wordnet-ova za različite jezike (slika 1),
- podsistem za generisanje klasa složenih reči na osnovu rezultata morfološke analize njenih komponenata,
- podsistem za interakcije sistema elektronskih rečnika i ontologija,
- podsistem za konverziju iz jednog kodnog rasporeda u drugi, konverzije iz jednog formata resursa u drugi, kao i konverzije lokalnih gramatika,
- okruženje za izgradnju i eksploataciju paralelizovanih korpusa, uključujući i vizuelizaciju u HTML-u,
- integrisano okruženje koje objedinjuje kompleksne aplikacije (Intex, NooJ, Unitex, Visdic), jezičke resurse i podsistem za ekspanziju upita,
- korišćenje i prezentacija paralelizovanih tekstova,
- veb servis za ekspanziju upita i veb aplikacija za prosleđivanje generisanog upita Google mašini za pretraživanje korišćenjem Google AJAX API-a.

Mada je WS4LR uglavnom korišćen za srpski jezik, njegova upotreba nije zavisna od jezika. Jedina pretpostavka je da resursi postoje ili da se razvijaju prema predviđenim formatima i metodologijama.

Raznovrsnost kriterijuma za postavljanje upita koje ovaj sistem omogućava rezultat je objedinjavanja skoro svih raspoloživih resursa, tako da su velike mogućnosti koje ovaj softverski alat pruža prilagođene različitim profilima korisnika. Naime, pretraživanje tekstova je moguće po sledećim kriterijumima:

- jednostavno sravnjivanje niski karaktera,
- lema sa svim svojim flektivnim oblicima, tj. morfološko proširenje zadate leme,
- koncept, gde se na osnovu zadate leme bira u semantičkom resursu odgovarajući koncept i literali pomoću kojih se taj koncept leksikalizuje,
- pretraživanje po konceptima uz morfološko proširenje,
- pretraživanje po konceptima, prošireno na drugi jezik.



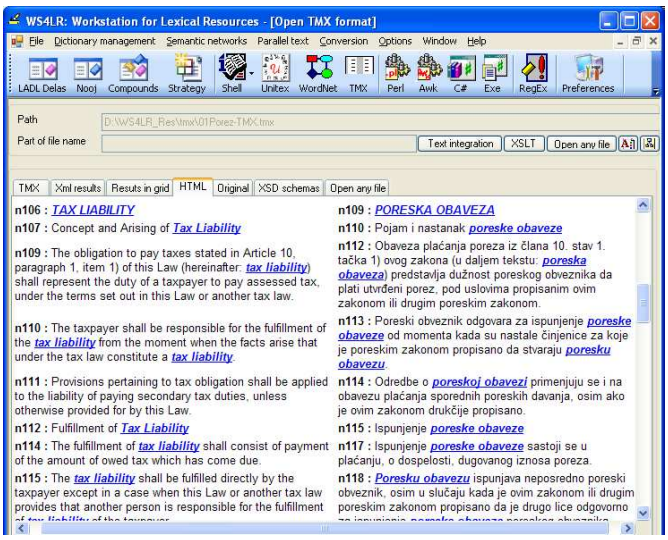
Slika 2. Proširivanje upita integrisanjem više resursa

Jedan od ciljeva WS4LR-a je da omogući što bolje iskorišćenje paralelizovanih tekstova koji predstavljaju resurs velikih mogućnosti, a koji zahteva veoma mnogo napora i znanja da bi se konstruisao. Objedinjavanje raspoloživih resursa se ilustruje pravljenjem konkordansi, odnosno izdvajanjem delova teksta koji zadovoljavaju određeni kriterijum.

Na slici 2. je prikazana forma softvera WS4LR za generisanje proširenog upita, u kom je navedena složena reč *'poreska obaveza'*. Korišćenjem morfoloških rečnika, lokalnih gramatika i sistema pravila su generisani svi oblici ove složene reči, a povezivanjem ovog koncepta putem WordNet-a je pronađen odgovarajući sinset u engleskom i pridružen je postavljenom upitu. Detaljniji opis ekspanzije upita će biti dat u narednom odeljku.

Ovako proširen upit je primenjen na paralelizovani tekst u TMX formatu da bi se izdvojile uparene rečenice koje odgovaraju postavljenom upitu. Na slici 3 je prikazan rezultat primene proširenog upita na jedan TMX[2] dokument.

[2] TMX 1.4b Specifacion, OSCAR Recommendation, 26 april 2005, http://www.lisa.org/tmx



Slika 3. Rezultat primene proširenog upita na TMX

4. WS4QE

Predstavljeni leksički resursi pružaju mogućnosti za sistematično rešavanje izloženih problema pri postavljanju upita, ali je pri tome je, u cilju postizanja ravnoteže između odziva i preciznosti, potrebno korisniku obezbediti što veću fleksibilnost u izboru niski od kojih će se formirati konačni upit. Ovaj problem je naročito izražen kod pretraživanja resursa na vebu, tako da je u tom cilju razvijena aplikacija WS4QE (Workstation for Query Expansion - Radna stanica za proširivanje upita) koja, zajedno sa pratećim veb servisima, omogućava da se niz zadataka sada reši i preko veba. Treba naglasiti da bi možda adekvatniji termin bio "podešavanje" upita. Naime, pored proširivanja skupa niski koje će biti uključene u upit, korisniku se pruža i mogućnost njihovog izbora, odnosno brisanja niski iz proširenog upita za koje korisnik proceni da mogu značajno umanjiti preciznost i time narušiti ravnotežu između odziva i preciznosti.

Raznovrsne mogućnosti za podešavanje upita koje omogućava WS4QE rezultat su, kao i u slučaju WS4LR, integracije više raspoloživih resursa. WS4QE, kao i WS4LR, daje korisniku mogućnost da upit proširi morfološki, semantički ali i na još jedan jezik (engleski). Sem toga, WS4QE daje korisniku još šire mogućnosti kontrole nad formiranjem upita, jer sem proširivanja, omogućava i njegovo sužavanje.

Morfološko proširenje zasniva se na korišćenju morfoloških rečnika prostih reči i složenica. Uz svaku od varijanti postavljanja upita korisniku se nudi mogućnost morfološkog proširenja, prostim izborom odgovarajućeg polja. U tom slučaju WS4QE za zadatu ključnu reč pronalazi uz pomoć srpskog morfološkog rečnika sve njene flektivne oblike, kako za proste reči, tako i za složenice, i formira složeni upit povezujući ih logičkim "ili" vezama. Korisnik na isti način, prostim izborom odgovarajućih polja, bira da li želi da upit postavi na ćiriličnom ili latiničnom pismu, ili na oba.

Za semantičko proširenje upita koristi se srpski wordnet, tako što za zadatu ključnu reč WS4QE izdvaja sve sinsetove u kojima se ta reč nalazi i nudi ih korisniku. Korisnik dobija uvid u sve koncepte na koje se ključna reč odnosi, i to kroz skupove sinonima koji se za te koncepte koriste, kao i

definiciju samih koncepata. Potom mu se pruža mogućnost da, ukoliko želi, obriše neke od ovih sinsetova ako zaključi da se oni pre odnose na koncepte koji za njega nisu od interesa. Upit se može i dodatno semantički proširivati izborom određene semantičke relacije, na primer, relacije nadređenosti i podređenosti, a u tom slučaju će se među odabranim sinsetovima, pored navedene osnovne grupe sinsetova, pojaviti i sinsetovi koji odgovaraju nadređenim i podređenim konceptima iz osnovne grupe.

Kada je završen izbor koncepata koji su od interesa, WS4QE iz njih generiše zajednički skup reči. I tu se korisniku nudi mogućnost da neke od tih reči isključi iz upita. Motivacija za isključivanje neke od odabranih reči može ležati u činjenici da je njena semantička relevantnost za koncept mala, a da ta reč pri tome može generisati veliki broj irelevantnih dokumenata jer joj je, kao višeznačnoj ili homonimnoj, semantička relevantnost za neki drugi koncept, koji nije od interesa, znatno veća. Podešavanjem upita, uz izbor koncepata i ključnih reči, može se značajno podići preciznost odgovora na upit.

Slika 4. Kombinovanje semantičkog i morfološkog proširenja upita

Na slici 4. ilustrovana je mogućnost sistema WS4QE da kombinuje semantičko i morfološko proširenje upita, uz podešavanje pisma. Na upit za ključnu reč "kuća"dobijena su četiri sinseta, sa ukupno tri različite ključne reči. Uz svaki sinset se nalazi definicija koncepta, kao i mogućnost njegovog brisanja ali i uvida u odgovarajuće drvo nadređenih i podređenih pojmova, što korisniku može pomoći da odluči o eventualnom daljem semantičkom proširivanju upita. Iz ta četiri sinseta generisana je lista od tri ključne reči ("literala") od kojih se svaki može eventualno i obrisati.

Kada se korisnik definitivno odlučio za ključne reči, može pristupiti morfološkom proširenju, i eventualnoj promeni pisma na kome je zadat osnovni upit ili dodavanju još jednog pisma. Na dnu ekrana vidi se deo proširenog upita koji se sastoji od ključnih reči ili niski dobijenih morfološkim proširenjem upita koji je prethodno semantički proširen, uz promenu pisma sa latinice na ćirilicu.

Leksički resursi, integrisani kroz veb servis WS4QE, su primenjeni i za proširenje upita u Geološkom informacionom sistemu Republike Srbije (GeolISS). GeolISS je razvijen kao prostorna baza podataka i predstavlja centralni repozitorijum za digitalno arhiviranje, pretraživanje, analizu i vizuelizaciju geoloških podataka, i njihovo intranet i internet publikovanje. Obzirom na veliku količinu tekstualnih podataka kojima se prostornim objektima pridružuju različite vrste opisa, i u GeolISS-u se pojavljuje problem postavljanja upita nad tekstualnim podacima. Početne analize efekata proširivanja upita pokazale su da se proširenjem upita korišćenjem veb servisa WS4QE i leksičkih resursa mogu dobiti znatno bolji rezultati (Stanković, 2008).

5. ZAKLJUČAK

Koncept proširivanja upita zasnovan na leksičkim resursima izložen u ovom radu može se iskoristiti za unapređenje postavljanja upita Sistemu naučnih, tehnoloških i poslovnih informacija. Naime, SNTPI se prilikom pretraživanja tekstuelnih resursa suočava sa tipičnim problemima koji nastaju kada su u pitanju jezici sa bogatim morfološkim sistemom kao što je to srpski jezik. Sem morfološkog proširenja upita koje rešava većinu ovih problema, od interesa može biti i semantičko proširenje, koje je takođe na raspolaganju.

Pristup integraciji raspoloživih resursa prikazan je kroz jedan softverski alat koji pruža mogućnost sihronizovanog korišćenja većine od njih. Osnovna struktura ovog softvera i mogućnosti koje on pruža za efikasno upravljanje resursima, kao i proširenje upita su takvi da omogućavaju njegovo uspešno prilagođavanje različitim namenama, pa samim tim otvaraju i mogućnosti njeogovog korišćenja u okviru SNTPI.

LITERATURA

[1] Vitas D., Pavlović-Lažetić G., Krstev C., Popović Lj., Obradović I. (2003): „Processing Serbian Written Texts: An Overview of Resources and Basic Tools", Proc. of the International Workshop on Balkan Language Resources and Tools, Thessaloniki, Greece, S. Piperidis, V. Karakaletsis (eds.), pp. 97-104.

[2] Tufiş, D. (ed.). (2004) Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology, Bucureşti, Publishing house of the Romanian academy.

[3] Vitas D., Krstev C. (2007), „Extending Serbian E dictionary by the Use of the Lexical Transducers", Proc. of the 6th and 7th INTEX/Nooj Workshop, S. Koeva, D. Maurel, M. Silberztein (eds.), Formaliser les langues avec l'ordinateur: De INTEX à NooJ, Presses Universitaires de Franche Compte, Paris, 2007.

[4] Fellbaum C. (ed.) (1998) WordNet: An Electronic Lexical Database, The MIT Press.

[5] Maurel D., Vitas D., Krstev S., Koeva S., (2007) „Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian", BULAG n°32, 2007.

[6] Krstev C., Stanković R., Vitas D., Obradović I., "WS4LR: A Workstation for Lexical Resources", Proc. of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 2006, pp. 1692-1697.

[7] Stanković R. (2008) „Improvement of geodatabase queries within GeolISS", Pregled Nacionalnog centra za digitalizaciju 12/2008, Matematički fakultet, Beograd pp.65-74.