Nezavisno-domenski sistemi za izdvajanje imenovanih entiteta sa Veba bez ljudskog nadzora

Dragana Dudić

***Sadržaj* — Budućnost današnjeg Veba je Semantički Veb. Prelazak sa sadašnjeg Veba na Semantički Veb uključuje izdvajanje imenovanih entiteta sa današnjeg Veba u svrhu kreiranja RDF iskaza. S obzirom da količina informacija na Vebu eksponencijalno raste, za ovaj zadatak potreban je moćan nesupervizorski softverski sistem koji je nezavisan od domena. U ovom radu predstavljeno je i upoređeno pet postojećih sistema. Takođe, dat je predlog za nov sistem za izdvajanje entiteta.**

***Ključne reči* — izdvajanje imenovanih entiteta, izdvajanje informacija, Semantički Veb**

I. Uvod

Svakodnevno se Veb poveća za nekoliko hiljada milijardi Veb stranica. Samim tim, količina informacija na Vebu eksponencijalno raste iz dana u dan. Te novostvorene informacije su samo naizgled lako dostupne. Za njihovo pronalaženje u realnom vremenu neophodno je korišćenje alata za izdvajanje informacija.

Izdvajanje imenovanih entiteta je podzadatak jednog od osnovnih zadataka izdvajanja informacija (eng. Information Extraction) — prepoznavanja imenovanih entiteta. Kao takav je od velikog značaja za tehnologije u razvoju kao što je Semantički Veb. U temelju arhitekture Semantičkog Veba ističe se uticaj RDF-a, preciznije RDF iskaza, na više slojeve ove arhitekture. RDF iskaz je trojka koja se sastoji iz subjekta, predikata i objekta. Subjekat je

Dragana Dudić, Poljoprivredni fakultet u Beogradu, Nemanjina 6, 11080 Zemun, Srbija (e-mail: ddragana@agrif.bg.ac.rs)

resurs, objekat može biti resurs ili literal, a predikat predstavlja odnos između subjekta i objekta.

Za automatsko izdvajanje RDF iskaza iz Veba, od ključnog je značaja nesupervizorsko izdvajanje imenovanih entiteta. Od 1995. godine i MUC-6 konferencije, kada je započet rad sa imenovanim entitetima [1], pa do danas, razvijeni su mnogi alati za prepoznavanje imenovanih entiteta. Velika većina istih je zasnovana na ručno kreiranim pravilima i rečnicima vlastitih imena ili na tehnikama mašinskog učenja koje podrazumevaju kreiranje ručno obeleženog skupa dokumenata za učenje. Očigledno je da su ovi pristupi zavisni od ljudskog nadzora i domena. Alati zasnovani na ovakvim pristupima imaju vrlo visoke vrednosti za preciznost i odziv (95% - 98% za preciznost i 92% - 98% za odziv, prema [2]), gotovo koliko i čovek ali, bez obzira na to, imaju jako malu praktičnu vrednost za automatsko izdvajanje RDF trojki. Za rešavanje ovog problema mogu se koristiti nesupervizorski sistemi za izdvajanje imenovanih entiteta. Mnogi nesupervizorski sistemi su zavisni od domena što otežava njihovu efikasnu primenu u oblastima za koje nisu primarno napravljeni i stoga takvi sistemi neće biti razmatrani u ovom radu.

U ovom radu akcenat je na postojećim nesupervizorskim sistemima za izdvajanje imenovanih entiteta, koji su nezavisni od domena i oni su prikazani u poglavlju II. U odeljku III se upoređuju navedeni sistemi dok je u odeljku IV predložen dizajn novog nesupervizorskog sistema nezavisanog od domena. Dalje je predstavljen zaključak.

II. PREGLED SISTEMA

Izdvajanje imenovanih entiteta uključuje pronalaženje imenovanih entiteta, poput imena osoba, organizacija, lokacija ali i raznih vremenskih i numeričkih izraza. Čovek može vrlo jednostavno da izvrši ovaj zadatak. Ali obim posla, pogotovu ako je korpus za izdvajanje Veb, zahteva automatizaciju zadatka. Time zadatak značajno dobija na težini. Pritom se teži da sistem u što manjoj meri zahteva asistenciju čoveka jer nesupervizorski sistem može rekurzivno da se poziva i tako pronalazi veći broj entiteta. Minimizacija ljudskog udela u izdvajanju donekle umanjuje performanse sistema ali ne i njihov značaj.

Mnogi sistemi su prilagođeni izdvajanju iz specifičnih korpusa kao što su novinski članci vezani za određeni događaj i sudska dokumenta [3] – [5]. Uspešnost ovakvih sistema na drugim korpusima je više nego zanemarljiva. Otuda i potreba da, pored izuzimanja ljudskog nadzora, sistemi za izdvajanje entiteta sa Veba zahtevaju i nezavisnost od domena.

A. KnowItAll

KnowItAll [6] je nesupervizorski sistem za izdvajanje informacija, nezavisan od domena, koji izdvaja entitete na osnovu činjenice da je svaki imenovani entitet na Vebu redundantan. Ulaz za KnowItAll je klasa čije entitete treba izdvojiti kao i nekoliko generičkih ekstrakcionih šablona.

Proces izdvajanja entiteta započinje pozivanjem modula pod nazivom ekstraktor. Na osnovu zadatih generičkih ekstrakcionih šablona, ekstraktor šalje odgovarajući upit Veb pretraživaču. Dobijeni rezultati i ograničenja šablona određuju kandidate za entitete. Za svakog dobijenog kandidata procenjivač računa vrednost tačkaste uzajamne informacije (engl. Pointwise Mutual Information) između kandidata za entitet i fraza koje odgovaraju šablonima vezanim za klasu tog entiteta i to prema broju pogodaka pretraživača. Dobijene vrednosti tačkaste uzajamne informacije se prosleđuju naivnom Bajesovom klasifikatoru [7] koji dalje određuje da li je u pitanju entitet. Dalje se skup generičkih šablona za datu klasu proširuje uz pomoć algoritma za učenje šablona.

Učenje šablona počinje odabirom nekoliko već izdvojenih entiteta. Svaki od ovih entiteta predstavlja upit za Veb pretraživač. Za svako pojavljivanje entiteta u dobijenim rezultatima pretraživača, kreira se kandidat za šablon koji se sastoji od 4 reči pre entiteta, entiteta i 4 reči posle entiteta. Kandidati sa najboljim vrednostima za preciznost i odziv se biraju za nove šablone. Dalje proširivanje se vrši preko ekstraktora lista.

Ekstraktor lista je modul koji je funkcionalno sličan alatu Google Sets [8]. Ovaj modul najpre postavlja skup entiteta kao upit pretraživaču. Zatim se za svaku listu kreira omotač (engl. Wrapper) na osnovu koga se automatski izdvajaju novi entiteti.

B. WebKnox

WebKnox [9] je sistem za izdvajanje znanja sa Veba. Osnovni deo ovog sistema je ekstraktor entiteta koji se mahom zasniva na unpređenim idejama KnowItAll sistema. Sistem je nezavisan od domena i bez ljudskog nadzora je. Izdvajanje entiteta se vrši na osnovu nekoliko unapred zadatih koncepata tj. klasa. Tehnike koje autori ovog sistema koriste su: izdvajanje fraza, fokusirano pretraživanje i izdvajanje lista.

Tehnika izdvajanja fraza je vrlo slična tehnici koja se koristi u ekstraktor modulu iz KnowItAll sistema. Na osnovu nekoliko generičkih šablona se kreiraju odgovarajuće fraze koje se prosleđuju Veb pretraživaču. Svaka rezultujuća strana se ponovo pretražuje odgovarajućom frazom nakon čega se sve pronađene vlastite imenice izdvajaju kao entiteti.

Fokusirano pretraživanje za cilj ima izdvajanje entiteta iz lista. Postoji nekoliko generičkih šablona koji su usko povezani sa terminom lista. Na osnovu ovih šablona i zadatog koncepta se formiraju upiti za Veb pretraživač. Za svaki rezultat upita se proverava da li sadrži XPath putanju koja pokazuje na sve entitete u listi. Kada je najduža Xpath putanja pronađena, na osnovu nje se vrši izdvajanje entiteta iz liste. Da bi se prepoznale liste koje sadrže entitete za izdvajanje, WebKnox sistem koristi nekoliko heuristika.

Svrha tehnike izdvajanja lista je ista kao i tehnike za fokusirano pretraživanje. Način izdvajanja je sličan kao kod ekstraktora lista KnowItAll sistema, s tim što se koristi XPath omotač.

C. QL-Full

Log upita Veb pretraživača je tekstualni fajl koji se sastoji od niza zahteva pretraživača. Zahtev pretraživača sadrži informacije koje se odnose na upit ili na rezultat nekog upita i kao takav predstavlja bogat izvor imenovanih entiteta. QL-Full [10] je nezavisan od domena i potpuno nesupervizorski sistem koji vrši izdvajanje entiteta iz logova upita Veb pretraživača primenom heuristika i statističkih mera. Postupak izdvajanja entiteta započinje pronalaženjem kandidata za entitete iz logova upita Veb pretraživača. Zatim se na osnovu dva nivoa poverenja od kandidata za entitete biraju entiteti. Na kraju se eliminiše šum uz pomoć odgovarajućeg filtera.

QL-Full sistem se zasniva na pretpostavci je da ljudi često postavljaju kopirane delove nekog teksta kao upite. Na taj način se čuva kapitalizacija u upitima i to je upravo ono što se koristi za izdvajanje kandidata za entitete. Preciznije, kandidat za entitet se sastoji od svih reči upita koje počinju velikim slovom.

Za odabir entiteta iz skupa kandidata za entitete koriste se dva nivoa poverenja: nivo reprezentativnosti i novo samostalnosti. Nivoom reprezentativnosti se izražava verovatnoća da je odabrana reprezentacija entiteta ispravna. Nivo samostalnosti je zasnovan na pretpostavci da se kandidat za entitet često samostalno pojavljuje u logovima upita Veb pretraživača. Dobijeni skup entiteta se filtrira tako što se eliminišu svi kandidati za entitete koji u potpunosti sadrže nekog drugog kandidata za entitet.

D. SEAL

SEAL [11] je nesupervizorski sistem koji na osnovu tri unapred zadata entiteta neke klase pronalazi nove entitete te klase, bez obzira na domen. Sistem ima tri komponente: hvatač, ekstraktor i razvrstavač.

Hvatač preuzima Veb strane koje predstavljaju odgovor na upit dobijen konkatenacijom unapred zadatih entiteta. Strane koje je hvatač preuzeo se dalje šalju ekstraktoru. Funkcionisanje ekstraktora se zasniva na pretpostavci da se entiteti koji pripadaju istoj klasi pojavljuju u istom kontekstu na istoj Veb strani. Za svaku Veb stranu, ekstraktor kreira omotač na osnovu konteksta (prefiksa i sufiksa) u kom se javljaju početni entiteti. Uz pomoć omotača se izdvajaju kandidati za entitete.

Za određivanje konačnog skupa entiteta, koristi se razvrstavač. Razvrstavač najpre kreira graf u kome su čvorovi početni entiteti, omotači i kandidati za entitete. Kandidati se rangiraju prema sličnosti sa zadatim entitetima a sličnost se određuje uz pomoć slučajnog lutanja na grafu.

E. SEISA

SEISA [12] je nezavisan od domena i nesupervizorski sistem za proširivanje skupa entiteta, sličan SEAL sistemu. Za pronalaženje novih entiteta se koriste dva Veb izvora: HTML liste i logovi upita Veb pretraživača.

HTML liste se izdvajaju uz pomoć Veb popisivača (engl. Web Crowler). Na osnovu izdvojenih lista se kreira bezteżinski, bipartitni graf za liste i kandidate za entitete.

SEISA sistem za svaki log upita pronalazi entitet i kontekst za taj entitet koji se sastoji od najviše dva prefiksna ili sufiksna tokena. Kod izdvajanja entiteta iz logova upita Veb pretraživača, kreira se težinski bipartitni graf za entitete i kontekste a za težinu ivice se uzima vrednost zajedničke informacije (engl. Mutual Information) između entiteta i konteksta.

Za izdvajanje entiteta, iz bilo kog od navedenih Veb izvora, koriste se dva iterativna algoritma, jedan statički a jedan dinamički. Oba algoritma za ulaz imaju početni skup entiteta i bipartitni graf. Na osnovu mera relevantnosti i koherentnosti, od kojih se obe zasnivaju na proizvoljnoj poznatoj meri sličnosti (kosinusna, Žakardova,…), algoritam određuje konačan skup entiteta.

III. KOMPARATIVNA ANALIZA SISTEMA

Upoređivanje sistema zasnovanih na ručno kreiranim pravilima i rečnicima se vrši najpre na osnovu kvaliteta tih pravila i rečnika a potom i prema vrednostima mera kao što su preciznost i odziv. Sa druge strane, sistemi zasnovani na tehnikama mašinskog učenja su znatno složeniji. Stoga se pri upoređivanju ovakvih sistema zahteva upoređivanje tehničkih specifikacija, kao i performansi sistema koje se određuju uz pomoć odgovarajućeg odnosa

između preciznosti i odziva.

Nesupervizorski sistemi nezavisni od domena, koji služe za izdvajanje entiteta iz Veb izvora, su vrlo specifični za poređenje. Iako se ovi sistemi u nekoj meri oslanjaju na tehnike mašinskog učenja, ne možemo porediti njihove performanse na osnovu postojećih metrika, jer su u pitanju sistemi nezavisni od domena. Ako su vrednosti mera lošije za neku klasu, ne znači da neće biti odlične za neku drugu klasu. Rešenje bi moglo biti u poređenju prosečnih vrednosti mera za nekoliko klasa, ali se onda javlja problem odabira relevantnih klasa. Stoga, u ovom radu neće biti razmatrane performanse izražene preko mera. Evaluacija svakog od navedenih sistema je izvršena na osnovu specifičnih karakteristika (Tabela 1).

TABELA 1: POREĐENJE SISTEMA

Sistem	Jezička nezavisnost	Vrste entiteta	Veb izvori	Nivo izdvajanja	Nivo nadzornosti
KnowItAll	Ne	nenumerički	HTML tekst	Veb strana	6 šablona
WebKnox	Ne	numerički nenumerički	HTML	Veb strana	8 šablona
QL-Full	Da	nenumerički	logovi upita	-	-
SEAL	Da	numerički nenumerički	HTML	Veb strana	3 entiteta
SEISA	Da	numerički nenumerički	HTML logovi upita	Veb sajt	4 entiteta

A. Jezička nezavisnost

Nije redak slučaj da se sistemi za izdvajanje entiteta kreiraju ciljano za jedan jezik i to najčešće engleski. Engleski jezik je morfološki jednostavan jezik, stoga nije naročito komplikovano napisati ili, na osnovu anotiranog korpusa, naučiti pravila za izdvajanje entiteta. Zato su sistemi za izdvajanje podređeni ovom jeziku, najčešće supervizorski ili polu-supervizorski. Takođe, neosporno je da je najviše informacija na Vebu na engleskom jeziku. Ipak, broj Veb izvora koji nisu na engleskom jeziku svakodnevno raste i stoga ne treba zanemariti druge jezike kao vredne izvore novih entiteta.

KnowItAll i WebKnox sistemi koriste generičke šablone namenjene izdvajanju iz engleskog jezika. Da bi ovi sistemi bili nezavisni od jezika,

neophodno je dodati šablone za više jezika čime bi ozbiljno bila narušena nezavisnost od ljudskog nadzora ovih sistema. QL-Full je jezički nezavisan sistem, ali kvalitet izdvajanja zavisi od količine informacija u logovima upita koje su na određenom jeziku. Zato je kvalitet izdvajanja znatno manji kod manje popularnih jezika. SEAL je potpuno jezički nezavisan jer su Veb liste gotovo jednako zastupljene među mnogim jezicima. SEISA je takođe jezički nezavisan sistem. Za razliku od QL-Full sistema, SEISA sistem izdvaja entitete i iz lista, čime je povećan kvalitet izdvajanja za ostale jezike.

B. Prepoznavanje vrsta entiteta

Prema MUC-6 konferenciji [13], postoji 7 tipova imenovanih entiteta: osoba, organizacija, lokacija, datum, vreme, novac, procenat. U ovom radu će biti korišćena generalizacija ove tipizacije na numeričke i nenumeričke entitete.

Svi navedeni sistemi izdvajaju nenumeričke entitete, a SEAL i SEISA vrše i izdvajanje numeričkih entiteta. WebKnox sistem može da izdvaja numeričke vrednosti kao entitete ali se one češće izdvajaju kao vrednosti nekog atributa zadatog entiteta. Iako pod rečima podrazumeva alfanumeričke niske, QL-Full nije namenjen izdvajanju numeričkih entiteta. KnowItAll sistem ne podržava izdvajanje numeričkih entiteta jer tačkasta zajednička informacija nije pogodna za validaciju numeričkih entiteta.

C. Tipovi korišćenih Veb izvora

Veb je skup raznorodnih dokumenata. Stoga, izdvajanje entiteta sa Veba treba da obuhvati izdvajanje iz više vrsta dokumenata. Uopšteno, dokumente na Vebu možemo podeliti na HTML dokumente, opšte tekstualne dokumente (Word, PDF, PS,...) i struktuirane tekststualne dokumente kao što su logovi upita Veb pretraživača. S obzirom da je u ovom radu naglasak na Veb izvorima, opšti tekstualni dokumenti će biti zanemareni.

KnowItAll sistem je kreiran da bi izdvajao entitete iz HTML dokumenata ali način izdvajanja omogućava prilagođavanje izdvajanju i iz drugih vrsta dokumenata. SEISA sistem objedinjuje izdvajanje iz logova upita i HTML lista, koje su sastavni deo HTML dokumenata. WebKnox i SEAL maksimiziraju izdvajanje iz HTML dokumenata, dok su ostali izvori zanemareni. Slično je i sa QL-Full sistemom, gde je akcenat na izdvajanju iz logova upita Veb pretraživača.

D. Nivoi izdvajanja

HTML liste predstavljaju bogat izvor entiteta i često se koriste u sistemima za izdvajanje entiteta. Od navedenih sistema, samo QL-Full ne koristi HTML

liste.

Neretko se liste na Vebu protežu i na nekoliko Veb strana. Zato je važno da sistemi koji izdvajaju entitete iz lista vrše izdvajanje na nivou Veb sajta a ne Veb strane. Svaki sistem za izdvajanje entiteta koji koristi omotače ograničen je na izdvajanje na nivou Veb strane jer jedan omotač odgovara jednoj Veb strani. Iz tog razloga KnowItAll, WebKnox i SEAL vrše izdvajanje na nivou Veb strane. Sa druge strane, SEISA sistem izdvaja liste uz pomoć Veb popisivača koji vrše izdvajanje na nivou Veb sajta.

E. Nivoi ljudskog nadzora

Nesupervizorski sistemi su mahom intuitivno definisani. Oni teže nezavisnosti od ljudskog nadzora ali dozvoljavaju zadavanje malog broja početnih vrednosti, pod uslovom da dalji postupak ne uključuje ljudsku asistenciju.

Od navedenih sistema, jedino je QL-Full u potpunosti bez ljudskog nadzora. U KnowItAll sistemu postoji 6 zadatih generičkih šablona, dok u WebKnox sistemu postoji 8 generičkih šablona i to 3 za izdvajanje na osnovu fraza, a 5 za izdvajanje iz lista. SEAL i SEISA zahtevaju bar 2 zadata entiteta. Optimalni rezultati za SEAL sistem se dobijaju kada su zadata 3 entiteta, a za SEISA sistem za zadata 4 početna entiteta.

IV. PREDLOG SISTEMA

Pored navedenih pristupa, za Semantički Veb su od izuzetnog značaja i pristupi koji se ne zasnivaju na ustaljenim tehnikama izdvajanja informacija. Kod ovakvih pristupa, željene informacije se dobavljaju iz poznatih struktuiranih i polustruktuiranih Veb izvora poput Vikipedije, WordNet-a, GeoNames-a, IMDb-a i drugih, [14] – [16]. Sistemi zasnovani na ovakvim pristupima često nisu nesupervizorski, niti su nezavisni od domena, stoga nisu detaljnije predstavljeni u ovom radu, ali se mogu iskoristiti za dalji rad.

Jedan od načina za kreiranje sistema sa boljim performansama mogao bi se zasnivati na ujedinavanju pristupa zasnovanih na klasičnim tehnikama izdvajanja informacija sa novijim tehnikama, koje se zasnivaju na struktuiranim i polustruktuiranim Veb izvorima. Na taj način bi bile umanjene posledice mašinskih grešaka karakterističnih za informaciono-ekstrakcioni pristup, ali i ljudske greške karakteristične za noviji pristup. Ideja je da se izdvajanje imenovanih entiteta vrši u dva koraka. U prvom koraku će se vršiti izdvajanje imenovanih entiteta iz Vikipedijinih lista. Naime, elementi odgovarajućih lista sa Vikipedije će se iskoristiti za moguće entitete.

Pronalaženje odgovarajućih lista će se vršiti direktnom pretragom Vikipedije uz razmatranje Vikipedijinih stranica za preusmeravanje. Postoji nekoliko vrsta lista na Vikipediji i za svaku vrstu bi bio iskorišćen odgovarajući algoritam ili heuristika za izdvajanje entiteta iz njih. U drugom koraku, novi kandidati za imenovane entitete će se izdvajati iz rezultata Veb pretraživača.

Odabir novih imenovanih entiteta iz skupa kandidata za imenovane entitete će se vršiti na osnovu odgovarajuće mere sličnosti.

Taj sistem će ispuniti svaki od pet navedenih kriterijuma poređenja nezavisno-domenskih i nesupervizorskih sistema za izdvajanje imenovanih entiteta. On će biti potpuno jezički nezavisan jer je Vikipedija višejezična enciklopedija. Izdvajanje numeričkih entiteta neće predstavljati problem jer Vikipedija sadrži razne numeričke liste. Za izdvajanje imenovanih entiteta će se koristiti Vikipedija i HTML dokumenti ali, u zavisnosti od performansi, sistem se može dopuniti još nekim Veb izvorom. Izdvajanje imenovanih entiteta novog sistema iz Vikipedijinih lista će biti na nivou Veb strane, ali to neće predstavljati problem jer svaka lista na Vikipediji pripada tačno jednoj strani. Na kraju, ovaj sistem će u potpunosti biti bez ljudskog nadzora jer ne zahteva zadavanje početnih vrednosti.

V. ZAKLJUČAK

U ovom radu dat je pregled i poređenje aktuelnih, nezavisnih od domena i nesupervizorskih sistema za izdvajanje imenovanih entiteta koji, kao ekstrakcioni korpus, imaju jedan ili više Veb izvora i dat je predlog za novi, napredniji sistem za izdvajanje imenovanih entiteta. Poređenje navedenih sistema je izvršeno na osnovu pet poželjnih karakteristika svakog sistema za izdvajanje imenovanih entiteta. Nijedan od navedenih sistema nije ispunio sve kriterijume poređenja što ostavlja prostora za dalji rad i razvijanje ideja na kojima su zasnovani ovi sistemi.

Kombinovanjem navedenih pristupa sa pristupima koji izdvajaju informacije iz poznatih struktuiranih i polustruktuiranih Veb izvora, mogu se dobiti vrlo moćni nesupervizorski i nezavisno-domenski sistemi za izdvajanje imenovanih entiteta sa Veba. Predlog za jedan takav sistem je naveden u ovom radu. Iako performanse sistema nisu razmatrane u ovom radu, vrlo je važno da sistem ima zadovoljavajuće vrednosti mera performansi za veći broj domena. Naravno, performanse predloženog sistema se ne mogu predvideti, stoga dalji rad treba da uključi i istraživanje drugih Veb izvora, kako polustruktuiranih i struktuiranih, tako i nestruktuiranih Veb izvora.

ZAHVALNICA

Zahvaljujem se profesoru Dušanu Tošiću na nesebičnoj pomoći i podršci.

LITERATURA

[1] D.Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification", in Lingvisticæ Investigationes, vol. 30(1), S. Sekine, E. Ranchhod, Eds., Amsterdam: John Benjamins, 2007, pp. 3–26.

[2] E. Marsh and D. Perzanowski. (1998). MUC-7 Evaluation of IE Technology: Overview of Results. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Available: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html .

[3] R. Weischedel, "BBN: Description of the PLUM System as Used for MUC-6", in *Proceedings of the Sixth Message Understanding Conference MUC-6*, San Francisco: Morgan Kaufmann Publishers, 1995, pp.55–70.

[4] G. Krupka, "SRA: description of the SRA system as used for MUC-6", in *Proceedings of the Sixth Message Understanding Conference MUC-6*, San Francisco: Morgan Kaufmann Publishers, 1995, pp. 221–235.

[5] M. Bruckschen, C. Northfleet, D. da Silva, P. Bridi, R. Granada, R. Vieira, P. Rao and T. Sander. (2010). Named Entity Recognition in the Legal Domain for Ontology Population, in *Proceedings of the Workshop on Semantic Processing of Legal Texts (SPLeT-2010)*, Available: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W23.pdf , pp. 16–21.

[6] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates, "Unsupervised named-entity extraction from the Web: An experimental study", in *Artificial Intelligence*, vol. 165(1), Amsterdam: Elsevier Publishing Co., 2005, pp. 91–134.

[7] C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2011, ch. 13.

[8] Google Labs (2002, May, 20 – 2011, September, 5). Google Sets. No longer available.

[9] D. Urbansky, M. Feldmann, J. Thom, A. Schill, "Entity Extraction from the Web with WebKnox", in *Advances in Intelligent Web Mastering - 2*, V. Snášel, P. Szczepaniak, A. Abraham, Eds., Heilderberg: Springer, 2010, pp. 209–218.

[10] A. Jain and M. Pennacchiotti, "Domain-Independent Entity Extraction from Web Search Query Logs", ", in *Proceedings of the 20th international conference companion on World Wide Web*, New York: ACM Press, 2011, pp. 63–64.

[11] R. Wang and W. Cohen, "Language-independent set expansion of named entities using the web", in *Proceedings of IEEE International Conference on Data Mining*, P. Perner, Ed., Heilderberg: Springer, 2007, pp. 342–350.

[12] Y. He and D. Xin, "SEISA: set expansion by iterative similarity aggregation", in *Proceedings of the 20th international conference companion on World Wide Web*, New York: ACM Press, 2011, pp. 427–436.

[13] R. Grishman and B. Sundheim, "Message Understanding Conference – 6: A Brief History", in *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen: Center for Sprogteknologi, 1996, pp. 466–471

[14] J. Hoffart, F. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo and G. Weikum, "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages", in *Proceedings of the 20th international conference companion on World Wide Web*, New York: ACM Press, 2011, pp. 229–232.

[15] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, "DBpedia: A nucleus for a web of open data", in *Lecture Notes in Computer Science*, vol. 4285, K. Aberer, K.-S. Choi,N. Noy, D. Allemang, K.-I. Lee, LJB Nixon, J. Golbeck, P. Mika, D.

Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux, Eds. Heilderberg: Springer, 2007, pp. 722–735.

[16] A. Philpot, E. H. Hovy and P. Pantel, *Ontology and the Lexicon*, Cambridge:Cambridge University Press, 2008, ch. The Omega Ontology.

ABSTRACT

Semantic Web is the future of the current Web. The transition from the current Web to the Semantic Web involves named entitiy extraction from today's Web in order to create RDF statements. Since the amount of information on the Web grows exponentially, this task requires a powerful unsupervised, doman-independent entitiy extraction system. In this paper we present and compar five existing systems. Also, proposal for the new entity extraction system is given.

Domain-Independent and Unsupervised Named Entity Extraction Web-Based Systems

Dragana Dudic