Приказ ЕУРОЛАН 2015 летње школе из Рачунарске лингвистике

Јелена Митровић
jmitrovic@gmail.com,
Универзитет у Београду
Филолошки факултет

РАД ПРИМЉЕН: 10. фебруар 2016.
РАД ПРИХВАЋЕН: 14. фебруар 2016.

ЕУРОЛАН летња школа, одржана од 13. до 25. јула 2015. године је била дванаеста у низу летњих школа које се сваке две године одржавају у Румунији. Теме ових летњих школа увек су веома актуелне, па је и ове године за тему одабрана област Рачунарске лингвистике и уопште управљања подацима и знањем – Лингвистички повезани отворени подаци (енг. Linguistic Linked Open Data – LLOD). Локација овогодишње школе био је прелепи град Сибиу, у срцу Трансилваније, окружен планинама и предивном природом.

Двадесет истакнутих предавача који важе за највеће стручњаке из области Рачунарске лингвистике, а свакако из области LLOD, који су умногоме и допринели њеном развоју, држали су интензивне курсеве током две недеље трајања ове летње школе. Преподневна су била намењена теоретском делу упознавања са новим концептима, док су поподнева углавном била испуњена практичним радом, туторијалима и вежбама.

Учесници су дошли из Европе, али и из Кине и Аустралије. Једна од предности овакве врсте стицања нових знања и унапређивања професионалних вештина јесте и могућност повезивања и дељења искустава са колегама из целог света – што је свакако изузетно важно! Ово је прва летња школа коју сам похађала и изузетно ми је пријало то што сам имала прилику да две недеље проведем са људима који деле моју страст према лингвистици и коришћењу рачунарских технологија за обраду природних језика.

У основи свих предавања и туторијала на овој летњој школи у некој мери се говорило о Семантичком вебу. Семантички веб је пројекат израде универзалног медијума за размену информација постављањем докумената са значењем које рачунар може да обради на вебу. Главни циљ концепта семантичког веба јесте семантичка интероперабилност веб извора, те постојање инфраструктуре за машинску интерпретацију и закључивање о садржајима на вебу.

Оквир за описивање ресурса (Resource Description Framework – RDF) је концепт који је настао у потрази за ефикаснијим решењима за проналажење информација и један је од стандарда Семантичког веба. Представља општи метод за концептуално описивање информација – семантичких веза између

електронских извора. Састоји се из уређених тројки (триплета): Субјекат – Предикат – Објекат где је Субјекат RDF URI референца извора који описујемо, Предикат је RDF URI референца, семантичка веза, а Објекат је RDF URI референца, сам meta datum.

Кључна технологија Семантичког веба је и SPARQL (изговара се као енглеска реч „sparkle“, а рекурзивни је акроним за SPARQL Protocol and RDF Query Language). Овај језик је развијен посебно за претраживање RDF база података и представља W3C стандард. Остали важни елементи семантичког веба су свакако XML (eXtensible Markup Language) који одређује структуру података (RDF/XML), онтологије, то јест модели представљања знања или скупови дефиниција неких концепата и релација које између њих постоје, OWL (Web Ontology Language) који се користи за објављивање и дељење онтологија. Све ове технологије кључне су за функционисање парадигме LLOD.

Повезани отворени подаци (енг. Linked Open Data – LOD), који су основ за LLOD, према основним принципима које је 2006. године поставио Tim Berners-Lee, јесу подаци који: 1) Користе URIs (јединствене идентификаторе извора) као имена ствари; 2) Користе HTTP URIs везе како би се та имена могла пронаћи 3) Пружају корисне информације помоћу RDF и SPARQL стандарда; 4) Садрже везе према другим URI ради откривања што више ствари; 5) Податке треба отворити за коришћење преко отворених лиценци.

Графички приказ LLOD облака настао је на иницијативу радне групе за отворену лингвистику (енг. Open Linguistics Working Group)[1] и ова група га уређује и одржава, а све то у склопу мреже отвореног знања – OKFN (Open Knowledge Foundation Network)[2].

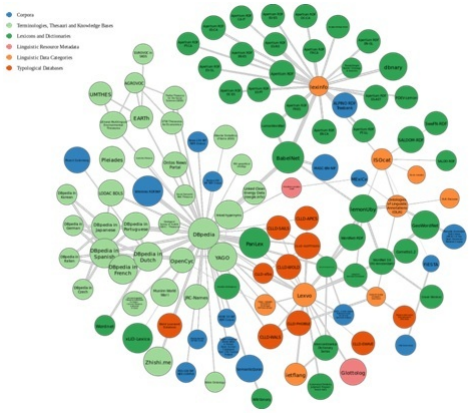
На слици 1 приказан је дијаграм на коме се види да су у LLOD облак укључени корпуси, базе знања, терминолошке базе, речници, лингвистичке категорије података, типолошке базе података.

DBpedia је веома важна у читавој LLOD и LOD парадигми. Она податке из Wikipedia страница трансформише у RDF. Садржи URI и друге метаподатке који се за сваку страницу формирају почевши од infobox делова Wikipedia страница. BabelNet је такође важан део LLOD облака. То је семантичка мрежа која обједињује WordNet, Open Multilingual WordNet (скуп свих отворених ворднетова), Wikipedia (највећа колаборативна енциклопедија), Wikidata (највећа колаборативна база знања), Wiktionary (највећи колаборативну речник), OmegaWiki (вишејезични речник средње величине).

Да би неки лексички ресурс био укључен у такозвани облак LLODпотребно је да буду испуњени следећи услови: 1) ресурс треба да буде доступан преко

[1] OWLG http://linguistics.okfn.org/2011/05/20/the-open-linguistics-working-group/

[2] OKFN https://okfn.org/



Слика 1. LLOD облак

разрешивих http:// (или https://) URI веза; 2) подаци које садржи морају бити разрешиви у RDF податке у неком од најпознатијих RDF формата (RDFa, RDF/XML, Turtle, N-Triples); 3) Мора садржавати бар 1000 уређених триплета (енг. triples); 4) мора бити повезан преко RDF веза са неким ресурсом који је већ у LLOD дијаграму, или мора имати најмање 50 веза према неким другим ресурсима; 5) Може му се приступити RDF претраживањем веба (енг. RDF crawling), преко RDF dump-а, или преко SPARQL сервиса који прима захтеве и враћа резултате (енг. SPARQL endpoint).

На овој летњој школи сазнала сам и да се недавно родила идеја ширег прихватања LLOD технологије у WordNet заједници, то јест LLOD-а као основног механизма за креирање веза између ворднетова на различитим језицима, и то преко интерлингвалног индекса (енг. ILI). Прихватање отворених лиценци и дељених формата довело је до много доступнијих података из светских ворднетова. Тако смо имали прилике да се упознамо са Универзалним ворднетом који за циљ има решавање проблема полисемије и синонимије, а све кроз вишејезичност, то јест повезивање сличних или истих концепата на различитим језицима.

Закључак приче о Повезаним лингвистичким отвореним подацима јесте да су они корисно решење за многе примене јер: 1) пружају интеграцију информација – могуће је на ефикасан начин пронаћи и искомбиновати информације из различитих извора; 2) омогућавају вишејезичност и мултилингвалне примене многих алата; 3) динамично објављивање – подаци на вебу нису статични, могу се видети њихове различите верзије и исправити

грешке; 4) помоћу модела заснованих на графовима могуће је приказати било који облик језичког ресурса; 5) Проналажење информација је структурирано, на пример, можемо добити одговор на питање „Која су имена свих добитника Нобелове награде пореклом из Француске?“.

Поред непроцењивих знања о начину на који фукционишу многе технологије семантичког веба и примера њихове практичне примене у конкретним пројектима, добила сам и много корисних предлога како бисмо српске лексичке ресурсе и алате могли да уврстимо у облак лингвистичких повезаних отворених података, и то од најбољих стручњака у тој области, од којих су неки поменуте технологије и осмислили, или значајно унапредили. Школа Еуролан је на мене свакако оставила веома позитиван утисак, а понајвише управо због инсистирања организатора да сви учесници проводе што више времена заједно, те су полазници курсева имали прилике да се боље упознају са предавачима и да од њих добију вредне савете. Следећа у низу ових летњих школа биће организована 2017. године и свакоме ко се бави Рачунарском лингвистиком бих топло препоручила да је похађа, а можда ћу то опет бити ја!