Четврти хакатон великих података о лингвистичким повезаним отвореним подацима

УДК 81'322.2:004.822

Тијана Радовић
tijana.n.radovic@gmail.com
Универзитет у Београду
Београд, Србија

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

САЖЕТАК: Четврти летњи хакатон великих података посвећен лингвистичким повезаним отвореним подацима (*4th Summer Datathon on Linguistic Linked Open Data, SD-LLOD-22*) у оквиру COST акције *NexusLinguarum* одржан је у Шпанији у мају 2022. Школа је окупљала заинтересоване истраживаче, професоре, студенте који су желели да стекну или прошире своја знања из области науке о лингвистичким повезаним подацима. Током школе представљен је спектар тема из области повезаних података, од онтологија, преко интеграције докумената, анотације и алата за обраду текста на природном језику заснованих на повезаним подацима и RDF-ом, до смерница за генерисање RDF-а и објављивање језичких ресурса као и давања увида у значај коришћења повезаних податак у лексикографији и терминологији. Организатор школе био је Политехнички Универзитет у Мадриду а школа је окупила преко 50 учесника како из академских институција тако и из индустрије.

КЉУЧНЕ РЕЧИ: лингвистички повезани отворени подаци, анализа сентимента, повезани подаци, RDF.

РАД ПРИМЉЕН: 28. новембар 2022.
РАД ПРИХВАЋЕН: 30. март 2023.

1. Увод

У оквиру COST акције *NexusLinguarum* одржан је Четврти летњи дан података о лингвистичким повезаним отвореним подацима (*4th*

Summer Datathon on Linguistic Linked Open Data, SD-LLOD-22), у Шпанији, месту Серседиљи код Мадрида у периоду од 30. маја до 3. јуна 2022. на Политехничком универзитету у Мадриду (*Universidad Politécnica de Madrid*). COST акцију *NexusLinguarum* (CA18209),[1] финансира Европска унија да би се остварила синергија између лингвиста, информатичара, терминолога, језичких стручњака и других заинтересованих истраживача широм Европе, и како би се истражила и проширила област науке о лингвистичким повезаним подацима (*linguistic linked data*) (Dojchinovski et al. 2021).

2. Програм и организација школе

Садржај школе SD-LLOD-22 састојао се од предавања из области повезаних података како би учесници добили теоријску основу и упознали се са предметом бављења и методологијом коришћеном у овој области. Потом, одржане су радионице у оквиру којих су се решавали конкретни проблеми из праксе из различитих области рада са подацима, са главним циљем да се учесницима из индустрије и академске заједнице пруже практична знања у области повезаних података примењених на лингвистику. И на крају, кроз практичан рад на реализацији мини пројеката су учесници радили у групама. Кроз сопствене мини пројекте примењивали су научено укључујући генерисање или коришћење лингвистичких повезаних података, а помоћ им је пружао један од неколико доступних ментора, стручњака за специфичне теме. Коначни циљ је био да се учесницима омогући да трансформишу постојеће лингвистичке податке и објаве их као повезане податке на вебу и да развију апликације које користе лингвистичке повезане податке. [2] Слични догађаји су и раније организовани: 2015. и 2017. у Серседиљи (Шпанија) и 2019. у Дагстулу (Немачка).

Различити језички ресурси попут речника, корпуса и база знања широко и отворено су доступни технологијом повезаних података. Међутим, и даље постоји велики број неповезаних ресурса који су похрањени у хетерогеним форматима и са различитим шемама представљања. Додатно, ти ресурси немају стандардне начине програмског приступа подацима (*Application Programming Interface*,

1. https://nexuslinguarum.eu/

2. Информације о догађају и линкови на ресурсе се могу наћи на званичном сајту https://datathon2022.linkeddata.es/

API). Представљање језичких ресурса као повезаних података има бројне предности као што су агрегација и интеграција језичких ресурса коришћењем заједничког модела података (*Resource description framework*, RDF), експлицитно повезивање, публиковање и претраживање на стандардизован начин (SPARQL), побољшано откривање скупова података и услуга, као и коришћење заједничких речника за представљање језичког садржаја и метаподатака. Укратко, повезани подаци омогућавају лакшу претрагу језичких ресурса, њихову доступност, интерпретабилност и поновну употребљивост (Gracia et al. 2022a).

Током школе, учесници су: 1) генерисали и објавили сопствене лингвистичке повезане податке из већ постојећих извора података, 2) примењивали принципе повезаних података и технологије семантичког веба у области језичких ресурса, 3) користили неке од најважнијих модела који се користе за представљање лингвистичких повезаних података, посебно *OntoLex-Lemon* (McCrae et al. 2017), 4) упознали се са токовима обраде текста на природном језику и апликацијама заснованим на повезаним подацима, 5) добили увид у потенцијалне предности повезаних података и могућности њихове примене за специфичне случајеве употребе.

Предавања у оквиру школе обрадила су следеће теме: онтологије и повезани подаци са акцентом на OntoLex-Lemon модел као модел лексикона за онтологије који подржава RDF (*Resource Description Framework*) и OWL (*Web Ontology Language*), две од три фундаменталне технологије коришћене у оквиру семантичког веба (Gracia et al. 2022b). Затим представљена је интеграција докумената, анотације и алати за обраду текста на природном језику засновани на повезаним подацима и RDF-у користећи Web Annotation и NIF (*NLP Interchange Format*) стандарде. Потом су представљене и смернице за генерисање RDF-а и објављивање језичких ресурса као и значај коришћења повезаних података у лексикографији и терминологији. На крају дат је осврт на целокупни значај употребе и примене лингвистички повезаних података.

3. *Sentimientos* – један од мини пројеката реализованих у току SD-LLOD-22

Као што смо претходно поменули, један аспект школе подразумевао је практични рад у групама кроз који су учесници кроз сопствене мини пројекте примењивали научено укључујући генерисање или коришћење

лингвистички повезаних података. Овај самостални практични рад замишљен је тако да учесници сходно својим интересовањима предлажу своје „мини пројекте“ који су укључивали конверзију у повезане податке постојећих скупова података, при чему су нарочито подстицани мини пројекти за језике са недовољно развијеним ресурсима. Представници Србије на овом догађају били су студенти докторских студија „Интелигентни системи“ при Универзитету у Београду: Тијана Радовић и Милош Кошпрдић. Заједно са колегом Мартином Алехандром Чајом са Универзитета у Мадриду и ментором Сином Ахмадијем са Универзитета у Болоњи радили су на пројекту *SD-LLOD-22 Sentimientos* чији је циљ био да обогати, конвертује и публикује лексикографске податке са информацијама о поларитету речи у формату повезаних података. За српски језик коришћен је лексикон *Senti-Pol-sr* (Stanković et al. 2022), а за немачки језик лексикон *PolArt* (Klenner, Fahrni, and Petrakis 2009). Структура оба лексикона је сасвим једноставна: лема и придружен поларитет исказан кроз колоне позитивно и негативно са дискретним вредностима 0 и 1.

Пројекат се састојао из три дела:

1) почевши од листе лема, прикупљени су лексикографски подаци из базе *Dbnary* (Sérasset 2012) користећи SPARQL упите. Креиран је шаблон онтологије: шема лексичких записа која се састојала од лексичког записа (*LexicalEntry*), врсте речи (*partOfSpeech*), леме (*lemma_label*), лексичког значења (*lexicalSense*) и поларитета (*hasPolarity*), што се може видети у коду који следи (плавом бојом су обележени делови који се мењају). Коришћењем представљеног обрасца је даље сваки запис из улазних речника трансформисан у RDF облик.

```
rdf_template = """
<:le_uri> a ontolex:lexicalEntry;
   lexinfo:partOfSpeech <:pos_uri>;
   rdfs:label 'lemma_label'@language_tag;
   ontolex:lexicalSense :senses_uris.

<:lemma_opinion> marl:describesObject <:le_uri>;
   marl:hasPolarity polarity.
"""
```

2) Да би се произвела такозвана *ttl* датотека, подаци су допуњени коришћењем SPARQL упита над лексиконом *Dbnary* како би се преузео лексички запис, врста речи и значење за конкретне речи из улазних

речника. У наставку приказујемо SPARQL упит за допуну постојећих података:

```
query = (
  "\n"
  " PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>\n"
  " PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>\n"
  " PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>\n"
  " \n"
  " SELECT ?le ?pos ?sense \n"
  " WHERE {\n"
  " ?le a ontolex:LexicalEntry .\n"
  " ?le rdfs:label \"xxxxx\"@<language2>.\n"
  " ?le lexinfo:partOfSpeech ?pos.\n"
  " ?le ontolex:sense ?sense\n"
  " }"
    "")
```

3) Упит се прослеђује приступној тачки *Dbnary*, у овом случају на следећи начин:

```
sparql = SPARQLWrapper( "http://kaiko.getalp.org/sparql")
sparql.setReturnFormat(JSON)

def run_query(word, language, language_):
    word_query = query.replace("xxxxx", word)
    word_query = word_query.replace("<language2>", language)
    sparql.setQuery(word_query)
    return sparql.queryAndConvert()["results"]["bindings"]
```

Пример лексичког записа који садржи лему, врсту речи и поларитет осећања би у овој трансформацији изгледао како је приказано на слици 1.

У овом мини пројекту коришћени су поменути српски и немачки лексикони сентимента и *Dbnary* али се приступ може проширити и на друге врсте лексикона.

У циљу интегрисања произведених алата у корисничку апликацију, произведени подаци су похрањени у базу *GraphDB*[3] и над њима је направљена мала апликација коришћењем програма Јава скрипт за претраживање похрањених података.

3. GraphDB

```
### word: hrana
<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>
  a ontolex:lexicalEntry;
  lexinfo:partOfSpeech <http://www.lexinfo.net/ontology/2.0/lexinfo#noun>;
  rdfs:label 'hrana'@sr;
  ontolex:lexicalSense
  <http://kaiko.getalp.org/dbnary/ell/__ws_1__srp__hrana__ουσιαστικό__1>.

<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1/opinion>
  marl:describesObject
  <http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>;
  marl:hasPolarity marl:positive.
```

Слика 1. Пример лексичког записа.

Коришћени и произведени подаци, као и развијени код су јавно доступни,[4] а додатна верзија је допуњена врстама речи из *Српског морфолошког речника* (Krstev 2008) и коришћењем лексичке базе *Лексимирка* (Stanković et al. 2018). Датотека *Senti-Pol-sr.ttl* доступна је за директан приступ и преузимање на страници Друштва за језичке ресурсе и технологије JePTex,[5] као и за претраживање коришћењем SPARQL језика на приступној тачки JePTex-а која је имплементирана коришћењем алата Fuseki.[6]

4. Закључак

Четврти летњи хакатон великих података о лингвистичким повезаним отвореним подацима нама као учесницима је био драгоцен начин за повезивање и размену знања, идеја и искустава са истраживачима сличних интересовања како из академских институција тако и из индустрије. Садржај је био разноврстан, са темама које су прилагођене знању полазника у различитим секцијама, тако да смо успели да проширимо раније стечена знања и добијемо смернице за будућа истраживања. Предавања теоријског типа помогла су нам да боље разумемо методологију и концепте повезивања, као и шеме

4. SD-LLOD-22 Sentimientos
5. http://llod.jerteh.rs
6. http://fuseki.jerteh.rs//dataset/Senti-Pol-sr/query

и стандарде репрезентације података. На радионицама, радом на конкретним проблемима упознали смо се са различитим технологијама у области повезаних података. Стицање знања и вештина током пет дана трајања школе је омогућило да генеришемо и објавимо сопствене лингвистичке повезане податке из већ постојећих извора података, што ће нам омогућити даља истраживања у овом правцу: генерисање сличних скупова података, надградњу кроз обогаћивање и проширивање публикованог лексикона. Такође, упознали смо се са токовима обраде текста на природном језику и апликацијама заснованим на повезаним подацима и уверили се у могућности њихове примене. Рад на мини пројекту *Sentimientos* допринео је обогаћењу ресурса на српском језику конвертовањем и публиковањем листе речи са придруженим информацијама о врсти речи и поларитетом у формату повезаних података (у виду *ttl* датотеке и на приступној тачки), а тај ресурс се може користити за будућа истраживања везана за одређивање сеинтимента текста, али и за објављивање лингвистичких података на сличан начин. Повезивање лексикона са примерима употребе који би били публиковани у *NIF* формату један је од првих задатака којим ћемо се позабавити у наредном периоду.

Захвалност

Представљени рад је подржала COST акција *CA18209 – NexusLinguarum "European network for Web-centred linguistic data science"*. Аутори се захваљују Милошу Кошпрдићу на учешћу у припреми скупа података.

Литература

Dojchinovski, Milan, Julia Bosque Gil, Jorge Gracia, and Ranka Stanković. 2021. "EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School." *Infotheca* 21 (1): 113–120. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.1.7.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022a. *SD-LLOD-22 Course Slides,* October. https://doi.org/10.5281/zenodo.7197674.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022b. *SD-LLOD-22 Materials,* October. https://doi.org/10.5281/zenodo.7197696.

Klenner, Manfred, Angela Fahrni, and Stefanos Petrakis. 2009. "Polart: A robust tool for sentiment analysis." In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009),* 235–238.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology, University of Belgrade.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Sérasset, Gilles. 2012. "Dbnary: Wiktionary as a LMF based Multilingual RDF network." In *Proc. of the 8th Int. Conf. LREC'12,* edited by Nicoletta Calzolari et al. ELRA. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data,* 31–38.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries – from File System to *lemon* Based Lexical Database." In *Proc. of the 11th Int. Conf. LREC 2018 - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), Miyazaki, Japan, May 7-12, 2018,* edited by John P. et al McCrae. ELRA. http://lrec-conf.org/workshops/lrec2018/W23/pdf/3_W23.pdf.