Limbajul natural ca mod de interogare a colecţiilor LinkedData

Anca Mărginean
Universitatea Tehnica Cluj Napoca
Baritiu 26-28, Cluj Napoca, Romania
anca.marginean@cs.utcluj.ro

Oana Marc
Universitatea Tehnica Cluj Napoca
Baritiu 26-28, Cluj Napoca, Romania
oana.marc@yahoo.com

REZUMAT
Chiar dacă Linked Data e un concept relativ nou, interogarea colecţiilor sale reînnoieşte vechile provocări ale interogării datelor într-o formă simpă dar suficient de puternică expresiv. Din perspectiva facilitării accesului la date în absenţa unei cunoaşteri prealabile a domeniului, cea mai firească abordare ar fi interogarea în limbaj natural. Acceptând limitările curente ale procesării limbajului natural, introducem primele rezultate obţinute în direcţia translatării interogării din limbaj natural în interogare SPARQL, translatare centrată pe analiza sintactică şi şabloane structurale generale combinate cu descrieri ontologice.

Cuvinte cheie
Web Semantic; parsare semantică; interogare SPARQL

Clasificare ACM
H.3.3 Information Search and Retrieval: Query formulation

INTRODUCERE

Linked Data înseamă utilizarea Web-ului pentru conectarea datelor care nu au fost conectate în prealabil în vederea facilitării accesului integrat şi uniform la date. Unul dintre cei mai mari jucatori ai acestui domeniul incă tânăr este *DBpedia*[1] - un efort al comunităţii de a extrage informaţii structurate din Wikipedia. Domeniul guvernamental sau medical vin să completeze plaja de domenii în care Linked Data a luat un avânt deosebit. Problema interogării acestor date devine astfel complicată nu doar datorită diferitelor formalisme de reprezentare a datelor ci şi datorită dimensiunii enorme atât a datelor cât şi a ontologiior de descriere.

Ne propunem să investigăm interogarea acestor date în limbaj natural, pornind de la premisa Web-ului Semantic de acceptare a limbajului RDF(Resource Description Framework) drept standard de reprezentare a datelor. În ciuda numeroaselor rezultate în domeniul prelucrării limbajului natural[5], [3], parsarea sintactica şi mai ales cea semantică ridică încă o multitudine de probleme care vor limita expresivitatea sistemului nostru. Sistemul propus construieşte interogări SPARQL (Protocol and RDF Query Language) din interogări în limbaj natural prin îmbinarea metodelor existente de parsare sintactică cu o formă proprie de parsare semantica. Domeniul ales în prezent pentru testare este cel al datelor medicale, şi mai precis proiectul Bio2RDF.

LINKED DATA ŞI PROIECTUL BIO2RDF

În contextul acceptării tot mai largi a conceptului de Linked Data (Date interconectate), proiectul Bio2RDF se remarcă printre alte proiecte ce incearcă integrarea colecţiilor bioinformatice [2]. Cu Bio2RDF, documente din bazele de date bioinformatice publice, precum Kegg (Kyoto Encyclopedia of Genes and Genomes), PDB (Protein Data Bank), MGI (Mouse Genome Informatics), HGNC (HUGO Gene Nomenclature) şi baze ale centrului NCBI (National Center for Biotechnology Information) sunt accesibile în formatul RDF printr-un URI (Identificator Uniform de Resursă) unic.

DrugBank este o componentă a proiectului Bio2RDF care oferă informaţii despre medicamente, conţinând în jur de 766,920 triplete şi 4,800 medicamente. În DrugBank, fiecare medicament este o resursă care are proprietăţi şiruri de caractere precum *toxicitatea, categoria, indicaţiile, mecanismul de acţiune, sinonim, descriere, absorbţie*. Un alt tip de proprietăţi sunt cele care au drept valori alte resurse în loc de şiruri de caractere, precum *ddi-interactor-in, patent, dosaj, target, tip, la fel ca*. Conectarea resurselor intra si inter vocabular se poate observa în exemplele din figura 1, unde o resursă din vocabularul bio2rdf din DrugBank este în relaţie cu o resursă din vocabularul Pharmacogenomics Knowledge Base.

În acest exemplu, se observă triplete ce descriu resursa asociată medicamentului cu eticheta Lepirudin: organisme afectate, absorbţie şi ruta. Dacă valorile primelor două proprietăţi sunt şiruri de caractere, în cazul ultimei proprietăţi, valoarea este o altă resursă a cărei proprietate $label$ descrie modul de administrare “Intravenous”.

ACCESUL LA DATE ÎN LINKED DATA

Dincolo de problemele de integrare, accesul la date din Linked Data este greoi datorită necesităţii cunoaşterii prealabile a ontologiilor sau a vocabularului utilizat. Ceea ce propunem în aceasta lucrare este interogarea acestor date în limbaj natural. Considerăm că o astfel de abordare este justificată în ciuda limitărilor procesării limbajului natural prin faptul că utilizatorul uman nu este străin de domeniul interogat şi se poate referi la entităţile domeniului în limbaj natural. Ceea ce nu cunoaşte utilizatorul uman şi considerăm că nici nu ar trebui să cunoască, sunt ontologiile utilizate pentru reprezentarea domeniului. Ontologia ar trebui să fie destinată maşinii, nu omului, iar modul în care este reprezentată semantica datelor ar trebui să fie cât mai transparent utilizatorului uman.

Structura şi funcţionalităţile sistemului

Figura 2 prezintă principalele module ale sistemului. Modulul $Parsare$ determină arborele de parsare pentru interogarea furnizată de utilizator în limbaj natural în limba

```
http://bio2rdf.org/page/drugbank:DB00001 rdfs:label "Lepirudin"
http://bio2rdf.org/page/drugbank:DB00001 :affected-organism "Humans and other mammals"
http://bio2rdf.org/page/drugbank:DB00001 :absorption "Bioavailability is 100% following injection"

http://bio2rdf.org/drugbank_resource:13e8d5e28fb251ec51bdad66d6544621
       :route http://bio2rdf.org/pharmgkb_vocabulary:361ee98c3d82f85e8095179351912761

http://bio2rdf.org/pharmgkb_vocabulary:361ee98c3d82f85e8095179351912761 rdfs:label "Intravenous"

<!-- Un exemplu de interac\c{t}iune intre doua medicamente -->
http://bio2rdf.org/drugbank:DB00001 drugbank:ddi-interactor-in http://bio2rdf.org/drugbank_resource:DB00001_DB00374
http://bio2rdf.org/drugbank:DB00374 drugbank:ddi-interactor-in http://bio2rdf.org/drugbank_resource:DB00001_DB00374
http://bio2rdf.org/drugbank:DB00374 rdfs:label "Ginkgo biloba"
http://bio2rdf.org/drugbank_resource:DB00001_DB00374 rdfs:label "DDI between Lepirudin and
Treprostinil - The prostacyclin analogue, Treprostinil, increases the risk of bleeding when combined with
the anticoagulant, Lepirudin. Monitor for increased bleeding during concomitant therapy."
```

Figura 1: Exemple de descrieri din cadrul proiectului Bio2RDF

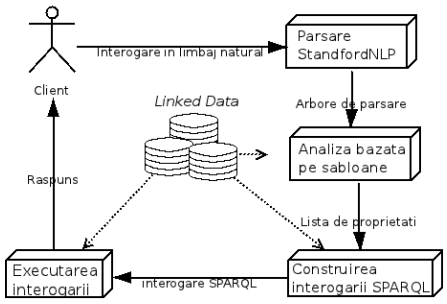


Figura 2: Modulele sistemului

engleza. În primul rând, interogarea este preprocesată în vederea evitării dependenţelor false, după care, folosindu-se API-ul Stanford Parser [6] , se realizează analiza morfologică şi sintactică a frazei, returnându-se arborele de parsare.

Modulul de $Analiza$ identifică tipul de întrebare prin suprapunerea arborelui peste un set de şabloane predefinite. Scopul acestor şabloane este de a extrage semantica asociată frazei şi de a identifica variabilele şi proprietăţile ce urmează a fi folosite în interogarea SPARQL. Lista de proprietăţi şi variabile astfel construită este preluată de către al treilea modul, cel de $Construire$ a interogarii SPARQL. In cele din urmă, interogarea generată este executată de către modulul de $Executare$ care se conecteaza la serviciul Bio2RDF de tip endpoint[1] care furnizeaza date despre medicamente din colectia DrugBank.

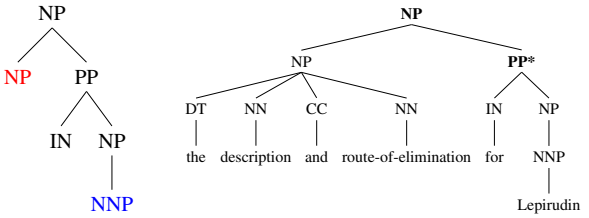
ANALIZA INTEROGĂRII BAZATĂ PE SABLOANE STRUCTURALE

Semantic interogării utilizatorului este extrasă pe baza unora şabloane structurale corelate cu ontologia curentă. Vom prezenta câteva dintre ele împreună cu interogarea SPARQL generată de sistem, urmând ca algoritmul de construire a interogării să îl detaliem în secţiunea următoare.

Rezultatul modulului *Analiză* este o listă de variabile şi proprietăţi, mai exact o **listă de tuple** de forma $\langle coloana$, $proprietate$, $inSelect$, $deCăutat\ \rangle$. Fiecărui aspect identificat în interogare îi corespunde o tuplă care în cele din urmă va deveni un pattern de triplet SPARQL. Componenta $Coloana$ reprezintă numele variabilei corespunzătoare unei proprietăţi cerute prin textul de intrare, *proprietate* este numele proprietăţii corespunzătoare în vocabularul Bio2RDF, în timp ce *deCăutat* poate avea valoare nulă sau numele unei anumite entităţi pe baza căreia se va face o filtrare. *inSelect* indică dacă o valoare a unei proprietăţi i) corespunde unui aspect solicitat de catre client, caz în care variabila dată de $coloana$ este inclusă în clauza $SELECT$ sau ii)este o proprietate auxiliară ce trebuie inclusă doar în clauza $WHERE$.

Reamintim că o interogare $SPARQL$ constă din două părţi: clauza $Select$ care identifică variabilele ce vor apărea în rezultat, respectiv clauza $Where$ care descrie graful şablon care restricţionează datele rezultat.

Şablonul 1



(a) Pattern 1 (b) What are the name and ...for Lepirudin

Figura 3: Şablonul 1 împreună cu un exemplu

Cel mai simplu şablon identificat este descris de structura din figura 3a. Interogările care respectă acest şablon sunt de forma **What is/are the** *list of properties* **for/of** *drug name* ? or **Find the** *list of properties* **for** *drug name*, unde proprietăţile pot fi binare sau n-are cu $n \geq 3$. Exemple de intrebari: i)*What is the substructure of Lymecycline?* sau ii)*Find the composition for Lymecycline.*, iii)*Find the description and toxicity for Lepirudin*, iv)*What is the description and interactor drugs for Lepirudin?*

Numele entităţii căutate este extras din structura **NNP** (proper noun phrase), în timp ce proprietăţile căutate sunt

[1] http://drugbank.bio2rdf.org/sparql

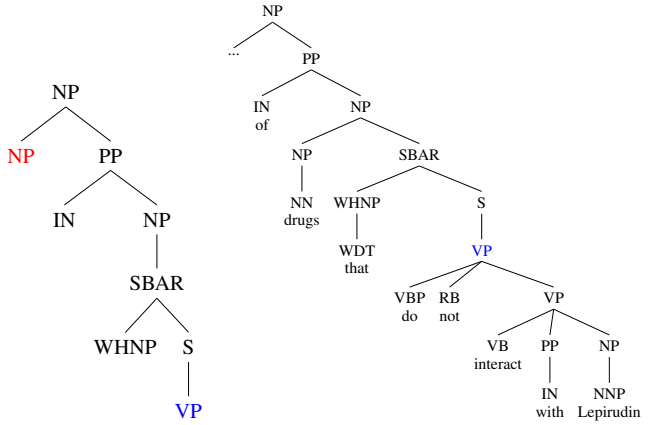
extrase din substantivele sau adjectivele identificate în subarborele structurii **NP** aflat în relaţie de determinare cu structura prepoziţională **PP**.

Pentru interogarea "What is the description and route of elimination for Lepirudin?" se obţine un arbore sintactic din care componenta relevantă este cea inclusa în figura 3b. Pe baza acestui arbore, modulul de *Analiza* construieşte următoarea listă de tuple din care ulterior modulul de *Construire* generează interogarea SPARQL.

coloana	proprietate	inSelect	deCăutat
description	⟨http://purl.org/dc/terms/description ⟩	true	nul
route-of-elimination	⟨http://bio2rdf.org/drugbank_vocabulary:route-of-elimination⟩	true	nul
name	⟨http://www.w3.org/2000/01/rdf-schema#label ⟩	false	Lepirudin

```
SELECT ?description ?routeofelimination
WHERE {
 ?a <http://www.w3.org/2000/01/rdf-schema#label> ?name.
 ?a <http://purl.org/dc/terms/description>
                        ?description.
 ?a <http://bio2rdf.org/drugbank_vocabulary:
      route-of-elimination> ?routeofelimination
  FILTER REGEX( str(?name), "Lepirudin", "i").
}
```

Şablonul 2



(a) Şablon 2 (b) What are the name and toxicity of drugs that do not interact with Lepirudin?

Figura 4: Şablonul 2 împreună cu un exemplu

Al doilea şablon structural (figura 4) permite ca proprietăţile solicitate să nu fie ale unui anumit medicament ci ale unei clase de medicamente. Această clasă poate fi determinată de

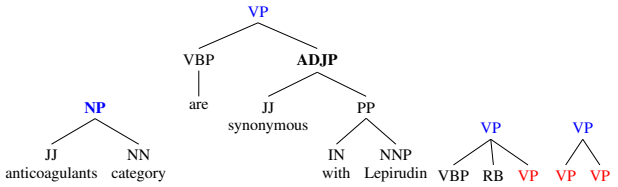
2.1 apartenenţa sau lipsa apartenenţei la o categorie, respectiv prezenţa interacţiunii: *drugs that are (not) în Anticoagulant category*

2.2 sinonimia cu alt medicament: *drugs that are synonymous with Lepirudin*

2.3 lipsa interacţiunii cu alt medicament: *(do not) interact with Lepirudin*

2.4 combinaţie ale primelor trei criterii: *drugs that interact with Ginkgo Biloba, are în anticoagulants category, do not interact with Lepirudin, and are synonymous with Hirudin?*

Şablonul 2 identifică structuri de genul **What are/is the** *list of properties* **for drugs that** *una dintre cele patru clase mai sus menţionate*. Similar şablonului 1, proprietăţile solicitate sunt identificabile pe baza substantivelor sau a adjectivelor din structura **NP**. Clasa medicamentelor este furnizată de subarborele structurii verbale **VP** pentru care în figura 5 se detaliaza câteva dintre tiparele posibile. Menţionăm că arborii includ cuvinte de instanţiere doar pentru o mai buna intelegere, şabloanele conţinând doar structuri sintactice sau morfologice, nu şi cuvinte concrete.



(a) 2.1 (b) 2.2 (c) 2.3 (d) 2.4

Figura 5: Clase de entitati - Sub-şabloane ale şablonului 2

Din motive de spaţiu nu vom prezenta celelalte şabloane utilizate în sistem, dar menţionăm câteva tipuri de întrebări acoperite de acestea:

- What type of drug is Lepirudin?
- What are the drugs from analgesics category ?
- Is there an interaction between Lepirudin, Thrombin, Hirudin?
- Is it safe/unsafe/dangerous to combine Lepirudin with Thrombin?

CONSTRUIREA INTEROGĂRII SPARQL

Odata extrase variabilele şi relaţiile dintre ele, modulul de *Construire* generează interogarea SPARQL.

Algorithm 1 Construirea interogarii SPARQL

```
1:  L ← analyse(arboreledeparsare)
2:  select ← "SELECT"
3:  where ← "WHERE{"
4:  for all el în L do
5:     if el.inSelect then
6:        column ← el.getColumn()
7:        if column contains "interact" then
8:           column ← "nameb"
9:        end if
10:       select+ ← "?" + column + " "
11:    end if
12: end for{Construirea clauzei Select }
```

Construirea interogarii SPARQL - continuare

13: **for all** el în L **do**
14: $column \leftarrow el.getColumn()$
15: $select+ \leftarrow$ "?a"$+el.getProperty+$"?"$+column+$"."
16: **if** $column$ $contains$ "$interact$" **then**
17: $select+ \leftarrow$ "?b" $+$ $el.getProperty$ $+$ "?" $+$ $column$ $+$ "."
18: $select+ \leftarrow$ "?b rdfs:label ?"+nameb+"."
19: filter (?name != ?nameb) ."
20: **end if**
21: **if** el.searchingFor **then**
22: $where+ \leftarrow$ "FILTER REGEX(str("+
23: "?"+$column$+"),"$+el.getSearchingFor+$",i)"
24: **end if**{REGEX for drug name}
25: **end for**
26: $where+ \leftarrow$ "}" {construirea clauzei Where}

Pentru construirea clauzei $select$ se parcurge lista de tuple L şi pentru fiecare tuplă cu componenta $inSelect$ setată la valoarea $true$ se adaugă în clauză o variabilă cu numele dat de componenta $coloana$ a tuplei. În cazul clauzei $Where$, pentru fiecare tuplă se adaugă câte un triplet format dintr-o variabila $?a$, proprietatea $property$ si variabila din $coloana$. Dacă componenta $deCautat$ a tuplei este $true$, se va adăuga o filtrare după proprietatea din tuplă.

Se poate remarca tratarea specială a proprietăţii $interact$, motivul fiind faptul că această proprietate nu există între două medicamente, ci între fiecare dintre cele două si o instanţă a conceptului *Drug-Drug-Interaction* (figura 1 - interacţiunea dintre Lepirudin şi Ginkgo Biloba). Pentru o întrebare de forma "What is the description and interactor drugs for Lepirudin?", rezultatul va fi

```
SELECT ?description ?nameb
WHERE {
 ?a <http://purl.org/dc/terms/description> ?description.
 ?a <http://bio2rdf.org/drugbank_vocabulary:
      ddi-interactor-in> ?interactor.
 ?b <http://bio2rdf.org/drugbank_vocabulary:
      ddi-interactor-in> ?interactor.
 ?b <http://www.w3.org/2000/01/rdf-schema#label> ?nameb.
 FILTER( ?name != ?nameb).
 ?a <http://www.w3.org/2000/01/rdf-schema#label> ?name.
 FILTER REGEX( str(?name), "Lepirudin", "i").
}
```

ALTE REZULTATE SI CONCLUZII

Potenţialul uriaş al colecţiilor din Linked Data a fost demonstrat şi de câştigarea premiului al treilea în cadrul competiţiei Semantic Web Challenge 2012 de către aplicaţia Open Self Medication[4]. În ce priveşte construirea interogărilor, menţionăm metoda descrisă în [8]. Este vorba de o metodă incrementală în care se porneşte de la cuvinte cheie urmând ca apoi, pe baza conceptelor din ontologia domeniului, utilizatorul să poate opta pentru anumite extinderi. În cazul nostru, ontologia nu e utilizată la nivelul utilizatorului, ci facilitează maparea pe un domeniu particular a unor metode generale, adică a şabloanelor structurale propuse.

QUEPY[2] este o aplicaţie în dezvoltare care încearcă interogarea datelor Web-ului Semantic în limbaj natural. Din ce cunoaştem noi, la momentul actual permite doar interogări simple, fără posibilitatea de a descrie clase de entităţi precum cele permise de Şablonul 2 al sistemului nostru. Direcţia de cercetare care a condus însă la abordarea propusă aici este cea a limbajelor de reprezentare a înţelesului (Meaning Representation Language) urmată în [3] sau [7]. În aceste două lucrări nu este vorba despre interogarea datelor RDF ci se prezintă metode de învăţare a parsării semantice sau a translatării către un MRL. Considerăm că în absenţa învăţării, nu se poate ajunge la un sistem de parsare semantică de un nivel acceptabil, aceasta fiind una dintre direcţiile următoare de lucru.

[2]http://quepy.machinalis.com/

În concluzie, s-au prezentat primele rezultate obţinute în direcţia interogării colecţiilor LinkedData în limbaj natural. Sistemul nostru are drept input un text în limbaj natural din care, pe baza analizei sintatice, a unor şabloane structurale şi a vocabularului de descriere a datelor, se obţine interogarea SPARQL. Şabloanele structurale facilitează extragerea semanticii interogării, de aceea considerăm că reprezinta cea mai importantă componentă a sistemului nostru. In mod cert, setul de şabloane trebuie extins, rafinat şi îmbogăţit cu metode de compunere şi de adaptare dinamică la ontologii.

Parte din cercetare s-a derulat in cadrul proiectului "Argumentare structurata pentru suport decizional cu constrangeri normative", programul PN-II Cooperari Bilaterale Romania-Republica Moldova, 2013-2014, UEFSCDI.

REFERINŢE

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC* (2007), 722–735.
2. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics 41*, 5 (2008), 706–716.
3. Chen, D. L., and Mooney, R. J. Learning to interpret natural language navigation instructions from observations. In *the Twenty-Fifth AAAI Conference on Artificial Intelligence, USA* (2011).
4. Cur, O. Open Self Medication on LOD. In *Proceedings of the Semantic Web Challenge co-located with ISWC2012* (Boston, US, November 2012).
5. Manning, C. D. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing - 12th International Conference, Japan* (2011), 171–189.
6. Marie-Catherine de Marneffe, B. M., and Manning, C. D. Generating typed dependency parses from phrase structure parses (2006).
7. Wong, Y. W., and Mooney, R. J. Learning for semantic parsing with statistical machine translation. In *HLT-NAACL* (2006).
8. Zenz, G., Zhou, X., Minack, E., Siberski, W., and Nejdl, W. From keywords to semantic queries - incremental query construction on the semantic web. *Journal Web Semantic 7*, 3 (2009), 166–176.