Natural Language Interfaces for SPARQL endpoints
- Hands-on tutorial on LODQA -

Jin-Dong Kim (DBCLS)
Agenda

- Intro to NLI SPARQL
- LODQA intro
- LODQA hands-on
- Related works
NLQA (Hybrid QA)

Ideal

Knowledge Bases

Language Processing

Query Generation

Rendering

Aggregation

Natural Language Query

Structured Query

Rendered Answer

Aggregated Answer

SPARQL Answer

SQL Answer

*query Answer

Linked (RDF) Data

RDB

Literature, Web, ...
NLQA (QA on LOD)

Language Processing

Query Generation

Natural Language Query

Structured Query

SPARQL

Linked (RDF) Data

Knowledge Bases

Rendered Answer

Aggregated Answer

Rendered Answer

Natural Language Query
Federated QA on LOD

Language Processing

Adapation to endpoints

Rendering

Aggregation

Natural Language Query

Pseudo SPARQL

SPARQL endpoints

SPARQL Answer

Rendered Answer

Aggregated Answer
Challenges

• Discrepancy
  ✔ Model representation (in NL)
  ✔ Data representation (in EP)
  ✔ Lexical discrepancy
  ✔ Structural discrepancy

which proteins phosphorylate IkB?

[Diagram showing relationships between protein, phosphorylation event, catalyzes, has_target, and IkappaB]
Typical approach

- Parsing
- Lexical Matching
- Structural Matching
Typical approach

Who wrote the Neverending Story?

Parsing

Lexical/structural matching

:Neverending_story

:has_author

:Michael.Ende

:who

wrote

the Neverending Story?
LODQA

- Open source project
- Highly portable to any SPARQL endpoint
  - Assumption: SPARQL endpoints in public are beyond anybody's control.
LODQA

- **Current state**
  - ✔ Project under progress
    - ➔ Focus on addressing **structural discrepancy** (⊙)
    - ➔ Lexical discrepancy (△)
    - ➔ Templating (×)
    - ➔ Relation matching is not yet implemented.
LODQA

• Current state
  ✔ Project under progress
    ➔ Incomplete system, but
    ➔ useful already to some extent.
  ✔ “not being perfect does not mean it's useless.”
  ✔ “will keep it useful during development.”
LODQA

• Three step approach
  1. Graphicator (parsing)
     ➔ Turns a natural language query into a pseudo graph pattern (PGP)
  2. Lexical mapping (dictionary lookup)
     ➔ To anchor the PGP on the target graph
     ➔ anchored PGP
  3. GraphFinder
     ➔ Search the KB graph for the anchored PGP.
LODQA (Linked Open Data Question Answering) is an open source project aiming at developing a system to generate SPARQL queries from natural language queries.

News

- (02/17-21/2014) OKBQA 2014 Hackathon is held in Jeju.
- (01/10/2014) Relation detection is improved to find the right pairs of terms in relation.
- (12/20/2013) LODQA is launched as an open source project.

Prototype Demo (targeting OMIM)

Input:
what side effects are associated with streptomycin?

Submit

Example

- what genes are associated with kabuki syndrome?
- what sign is associated with kabuki syndrome?
- what cellular dysfunction is associated with kabuki syndrome?
- what pathologic function is associated with kabuki syndrome?
- what neoplastic process is associated with kabuki syndrome?
- what anatomical abnormality is associated with kabuki syndrome?
- what genes are associated with alzheimer disease?
Query

what side effects are associated with streptomycin?

Semantic analysis (predicate-argument relation graph)
Query

What side effects are associated with streptomycin?

Semantic analysis (predicate-argument relation graph)

Enju HPSG parser
Query

what side effects are associated with streptomycin?

Semantic analysis (predicate-argument relation graph)

Pseudo Graph Pattern (PGP)

[side, effects]  [streptomycin]

[associated, with]
Query

what side effects are associated with streptomycin?

Semantic analysis (predicate-argument relation graph)

Step 1. Graphication

Pseudo Graph Pattern (PGP)

[side, effects] [streptomycin]

[associated, with]
Graph Pattern matching

Pseudo Graph Pattern (PGP)

[side, effects] [streptomycin]

[associated, with]

Target graph
Step 2. Lexical Mapping

• [side, effect]
  ✓ sider:side_effects
  ✓ sider:sideEffectName

• [streptomycin]
  ✓ drugbank:DB01082
  ✓ drugbank:DB00428
  ✓ Sider:5297
  ✓ sider:5300
Step 3. GraphFinder

Anchored PGP

sider:side_effects drugbank:DB01081

[associated, with]
Step 3. GraphFinder

Anchored PGP

sider:side_effects drugbank:DB01081

?p

Target graph
Final output: instances of the focused node

Anchored PGP

sider:side_effects drugbank:DB01081

?p

Target graph
Representational variations

Anchored PGP

sider:side_effects drugbank:DB01081

[associated, with]
Representational variations

Anchored PGP

sider:side_effects drugbank:DB01081

[associated, with]
Representational variations

Anchored PGP

sider:side_effects drugbank:DB01081

[associated, with]
Representational variations

Anchored PGP:

`sider:side_effects drugbank:DB01081`

[associated, with]
Operations for graph variation

① inversion

② split

③ join

④ instantiation
What proteins phosphorylate IkB?
What proteins phosphorylate IkB?
What proteins catalyze the phosphorylation of IkB?
④ Instantiation

What proteins catalyze the phosphorylation of IκB?
What proteins catalyze the phosphorylation of IkB?
The search space
Demo

- http://www.lodqa.org
Comparison to RelFinder

- **RelFinder**
  - ✔ [http://www.visualdataweb.org/relfinder.php](http://www.visualdataweb.org/relfinder.php)

- **GraphFinder generalizes RelFinder**
  - ✔ two instances → two, three, four, ...
  - ✔ → classes or instances
Summary

● Three step approach

1. Graphicator
   ➔ Turns a natural language query into a pseudo graph pattern

2. Lexical mapping
   ➔ To anchor the pseudo graph pattern on the target graph

3. GraphFinder
   ➔ Search the KB graph for the pseudo graph pattern
Summary

● Three step approach

1. Graphicator
   ➔ Turns a natural language query into a pseudo graph pattern

2. Lexical mapping
   ➔ To anchor the pseudo graph pattern on the target graph

3. GraphFinder
   ➔ Search the KB graph for the pseudo graph pattern

NLP task

LOD task
Summary

● Three step approach

1. Graphicator
   ➔ Turns a natural language query into a pseudo graph pattern

2. Lexicon
   ➔ To anchor the pseudo graph pattern on the target graph
   Representational difference needs to be absorbed

3. GraphFinder
   ➔ Search the KB graph for the pseudo graph pattern

NLP task

LOD task
Summary

- Three step approach

1. Graphicator
   - Turns a natural language query into a pseudo graph pattern

2. Lexical mapping
   - To anchor the pseudo graph pattern on the target graph

3. GraphFinder
   - Search the KB graph for the pseudo graph pattern

NLP task

LOD task

Representational difference needs to be absorbed

variation operations
Natural Language Interfaces for SPARQL endpoints
- Related Works -

Jin-Dong Kim (DBCLS)
Typical approach

• Parsing
• Lexical Matching
• Structural Matching
Typical approach

Who wrote the Neverending Story?

Parse:

- subj: who
- wrote
- obj: the Neverending Story?

Matching:

:has_author
:Michael_Ende

:Neverending_story
Frontiers

● NQ (2007)

● Aqualog (2007)
Frontiers

● ORAKEL (2007)

● QuestIO (2008)
Recent systems

- TBQA (AKSW, UManheim, ...)
  - Template-based SPARQL learner
  - http://linkedspending.aksw.org/tbsl/
- Treo (DERI)
  - 'direction' in Gallic
  - http://treo.deri.de
- LODQA (DBCLS, UCorado, ...)
  - Linked open data question-answering
  - http://www.lodqa.org
TBSL

• Parsing
  ✓ LTAG (lexical tree adjoining grammar)
  ➔ Tree transformation

• Lexical Matching
  ✓ ...

• Structural Matching
  ✓ Template generation
TBSL

● To address complex queries
  ✓ Who produced the most films?

● Generate templates
  ✓ SELECT ?y WHERE {
      ?x a onto:Film .
      ?x onto:producer ?y
  }
  ORDER BY DESC(COUNT(?x)) OFFSET 0 LIMIT 1
TBSL

• To address complex queries
  ✔ Who produced the most films?

• Generate templates
  ✔ SELECT ?y WHERE {
      ?x a onto:Film .
      ?x onto:producer ?y
  }
  ORDER BY DESC(COUNT(?x)) OFFSET 0 LIMIT 1
Treo

- Parsing
  - Dependency parsing
- Lexical Matching
  - Distributional semantics
- Structural Matching
  - ...
Treo

- **Lexical matching**
  - ✔ Distributional semantics
    - “linguistic items with similar distributions have similar meanings.”

Who is the daughter of Bill Clinton?

```
:Bill Clinton
  :child
  :religion
  :almaMaster
  ...
```
Lexical matching

Distributional semantics

“linguistic items with similar distributions have similar meanings.”

Who is the daughter of Bill Clinton?
LODQA

- Parsing
  - ✔ HPSG (Head-driven Phrasal Structure Grammar)
  - ➔ Graph transformation

- Lexical Matching
  - ✔ ...
  - ✔ Public sourcing lexical indexing

- Structural Matching
  - ✔ Graph variation operations
Future directions

- LODQA (DBCLS, UColorado, ...)
  ✓ Addresses Structural variation problem
- Treo (DERI)
  ✓ Addresses lexical variation problem
- TBQA (AKSW, UManheim, ...)
  ✓ Addresses quantifier modeling
Future directions

- **LODQA**
  - ✔ Invite contribution from the public
    - ➔ Open source
    - ➔ Open more information
    - ➔ Implement more open interface

Collaborations