Semantic Question Answering on Big Data

Tatiana Erekhinskaya

July, 2016
The Goal

Challenge:

• Find answers to complex questions in large structured and unstructured data resources

• Sample question: List Chinese researchers who worked with Kuznetsov, have publications on Zika virus and studied in US

Solution:

• Convert data into RDF storage

• Convert questions into SPARQL
Outline

• System Architecture
• NLP & Semantic Parsing
• RDF Representation
• Plain English Query to SPARQL
• Experiments & Results
• Use Cases & Future Work
System Architecture

Input Docs → Knowledge Extraction → Deep NLP → Question Processing → RDF Store → Semantic to RDF → Semantic to SPARQL → User Question
Natural Language Processing
Concept Extraction

- Hybrid approach combines machine learning classifiers, cascade of finite-state automata, and lexicons
- Uses existing medical ontologies: MeSH, SNOMED and UMLS Metathesaurus
- 80+ types of named entities: demographics, disease, symptom, dosage, severity, time course, onset, alleviating and aggravating factors
Semantic Parsing

• Extracts 26 predefined binary relation types: AGENT, THEME, LOCATION, TIME, etc.
• Maximum granularity, not limited to verb arguments: VALUE, PROPERTY, QUANTITY
• Robust basic representation, not for end users
Semantic Calculus

- Defines how and under what conditions a chain of relations can be combined into a high level custom relation

- Axioms: Possession(c1;c2) & ISA(c1, disease) & ISA(c2; organism) \(\Rightarrow\) HasDisease(c1; c2)

100 subjects with type 2 diabetes

QUANTITY

HAS_DISEASE

SEVERITY
RDF & SPARQL
RDF Representation

- 6.3 MB of text → 13 M triples, 1 GB of RDF XML
- Keep only relations of interest and tokens that participate in these relations
- For tokens: named entity type or is-event flag, lemma, synset, and reference sentence
Reasoning on the RDF Store

- OWLPrime
- SameAs: mentions
- Lexical chains: Wordnet-based relation sequence
Question Processing

- Full NLP & semantic parsing
- Expected answer type recognition (_human or organization, _date or _time, etc.)
- Answer type terms “which cartel”
- Maximum entropy model
SPARQL Query Formulation
Query Relaxation

• Synset relaxation: include hyponyms, parts, derivations

• On empty results: drop variable-description triples and semantic relations with little importance
Experiments & Results
Experimental Data

- Illicit Drugs domain
- 584 documents: Wikipedia + documents
- 6.3 MB of plain text
- 6,729,854 RDF triples
- 546 MB of RDF XML
Results: Question Answering

344 questions
Free text-search: 47% MRR
Semantic Approach: 66% MRR
Factoid: 85% MRR
Definition: 78% MRR
List: 68% MRR
Results: NL to SPARQL

34 manually annotated questions

- SELECT clauses: 85%
- WHERE clauses on triple level: 78%
- WHERE clauses on question level: 65%

Relaxation usage: 68% of queries
inSynset-relaxation sufficient for 31%
Error Analysis

73% caused by faulty or missing semantic relations

16% caused by query conversion: yes/no questions, and procedural questions
Conclusion

Use Cases
• Processing Pubmed for quality measures
• National Security: terrorism, law enforcement
• Foreign languages

Future Work
• Integration with LinkedData
• Rapid Customization