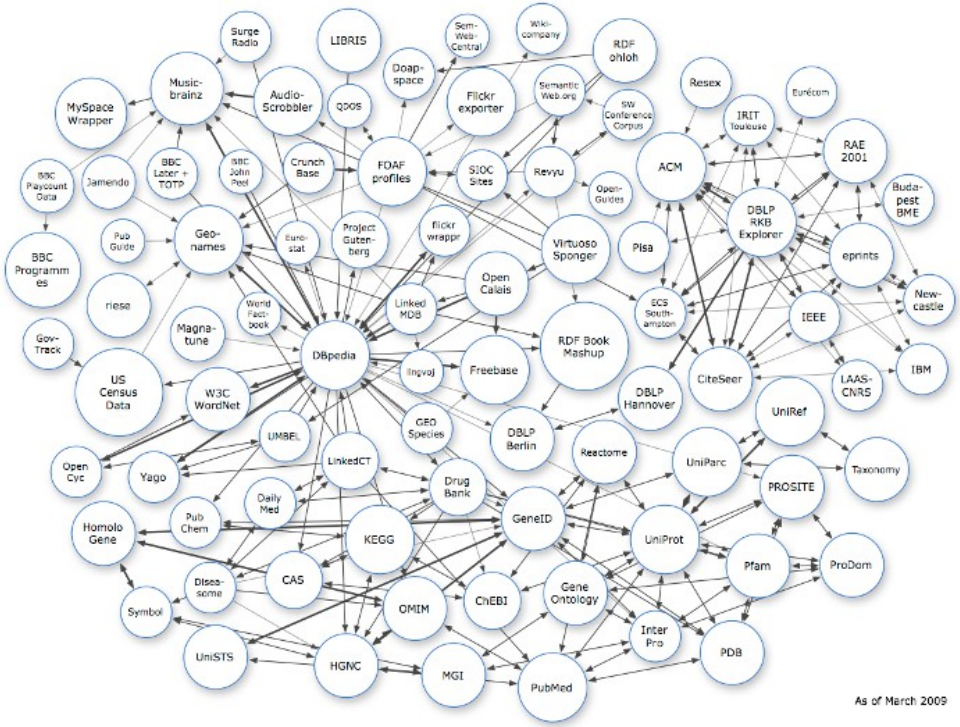
TECHNISCHE UNIVERSITÄT DARMSTADT

Template-based Question Answering over RDF Data



Quelle: http://www.alchemyapi.com/api/entity/ldata.png

Benjamin Lück

Resource Description Framework (RDF)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ursprünglich wurde RDF von der W3C für Metadaten entwickelt
- Ähnelt UML-Klassendiagrammen und ER-Modell
- Aussage bestehen aus Tripel / 3-Tupel
- Häufig als Tabelle dargestellt
- Gerichteter Graph, Tripel bilden RDF-Modell

Subjekt	Prädikat	Objekt
VW	produziert	Autos
VW	produziert	Passat
Mustermann	arbeitet bei	VW
VW	ist ein	Unternehmen
Passat	ist ein	Auto
Volkswagen	alternative	VW

- RDF grundlegender Baustein für semantisches Web
- Abfragesprachen: RQL, RDQL, SPARQL...

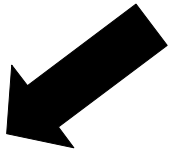
SPARQL Protocol und RDF Query Language (SPARQL)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Abfragesprache für RDF Datenbanken
- Standardisiert durch W3C
- Query Typen
  - SELECT
  - CONSTRUCT
  - ASK
  - DESCRIBE
- Operatoren

```
BASE <http://example.com/>
PREFIX abc: <exampleOntology#>
  SELECT ?capital ?country
  WHERE {
    ?x abc:cityname ?capital ;
       abc:isCapitalOf ?y .
    ?y abc:countryname ?country ;
       abc:isInContinent abc:Africa .
  }
```

Capital	Country
Algiers	Algeria
Luanda	Angola
Porto Novo	Benin
Djibouti	Djibouti
Cairo	Egypt



Problemstellung

TECHNISCHE UNIVERSITÄT DARMSTADT

- 7.4 Mrd. RDF Triplets
- Question answering systems: Aqua, NLP-Reduce, FREyA, PowerAqua
- Beispiele PowerAqua:
  - „Who wrote The Neverending Story?“

    <[person,organisation], wrote, Neverending Story>

    <Writer, IS_A, Person>

    <Writer, author, The Neverending Story>

Problemstellung

TECHNISCHE UNIVERSITÄT DARMSTADT

- „Which cities have more than three universities?“
  - PowerAqua: <[cities], more than, universities three>
  - Ziel:

```
SELECT ?y WHERE {
    ?x rdf:type onto:University . ?x onto:city ?y .
  }HAVING (COUNT(?x) > 3)
```

- „Who produced the most films?“
  - PowerAqua: <[person, organisation], produced, most films>
  - Ziel:

```
SELECT ?y WHERE {
    ?x rdf:type onto:Film .
    ?x onto:producer ?y .
  }ORDER BY DESC(COUNT(?x)) OFFSET 0 LIMIT 1
```

- Template:

```
SELECT ?x WHERE {
    ?x ?p ?y .
    ?y rdf:type ?c .
  }ORDER BY DESC(COUNT(?y)) OFFSET 0 LIMIT 1
```

Übersicht

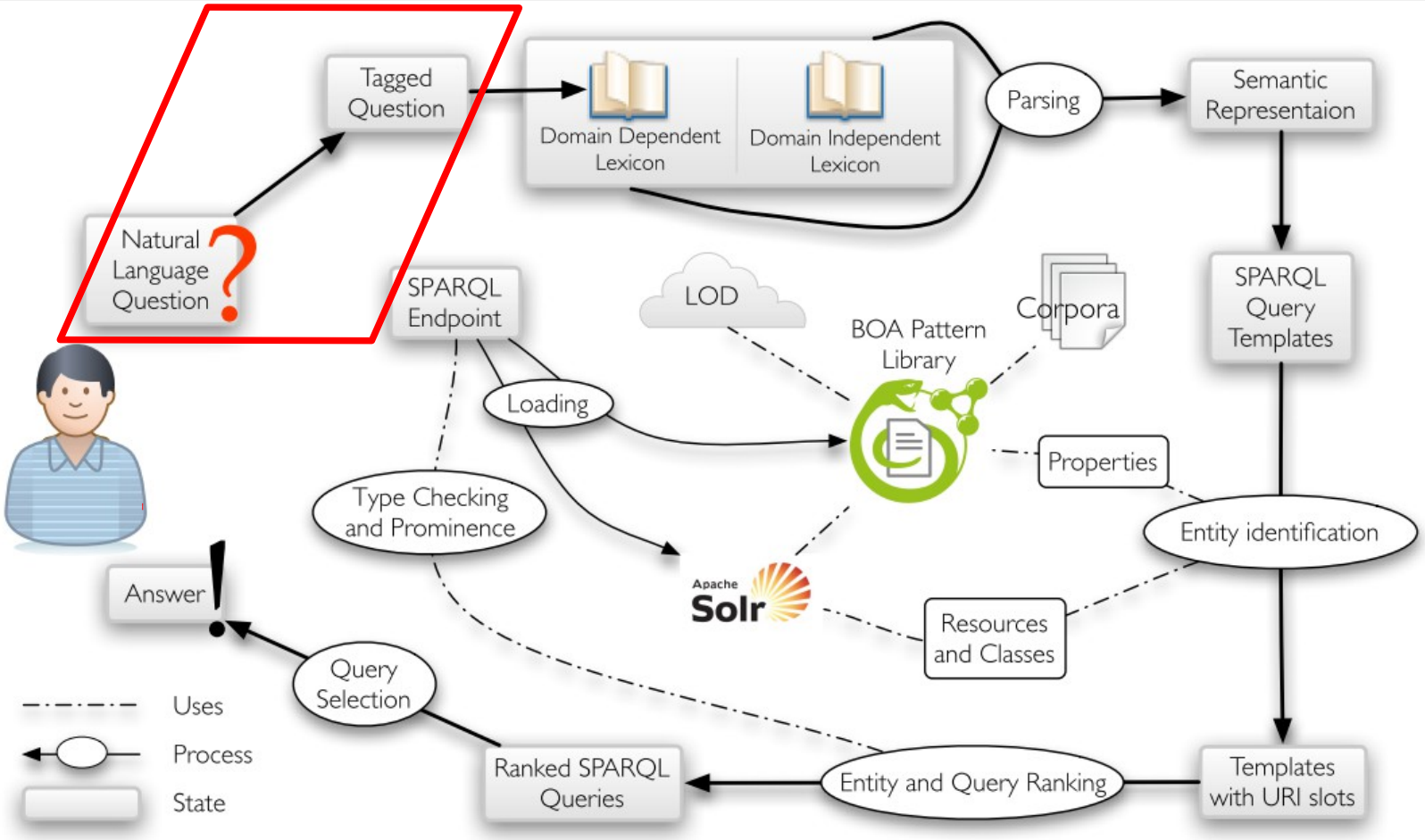


Figure 1: Overview of the template based SPARQL query generator.

POS Tagging

TECHNISCHE UNIVERSITÄT DARMSTADT

- Part-of-speech (POS)
- Identifizieren aller Wörter
  - Manuell
  - Computerlinguistik
- „Who produced the most films?“
  - → who/WP produced/VBD the/DT most/JJS films/NNS

Übersicht

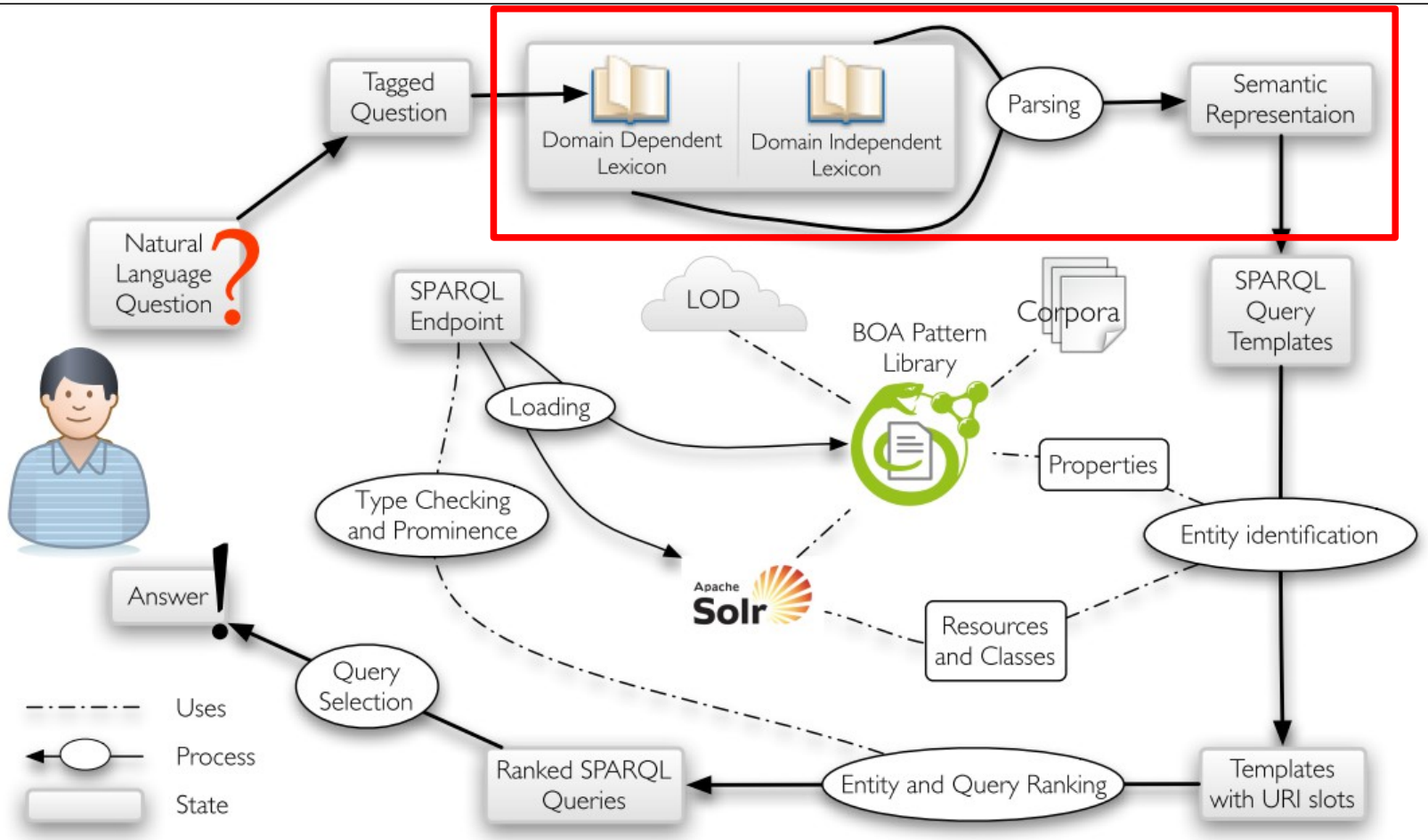


Figure 1: Overview of the template based SPARQL query generator.

Lexika und lexikalische Einträge

TECHNISCHE UNIVERSITÄT DARMSTADT

- Lexikon hat zwei Teile
  - Unabhängiger Teil

    Manuell erstellt, 107 Einträge
  - Abhängiger Teil

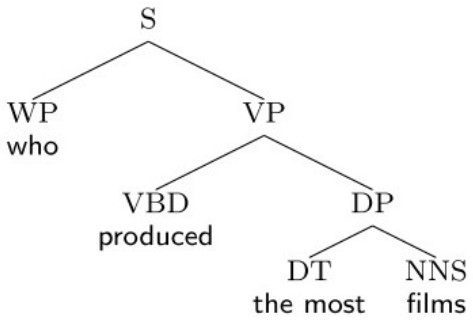
    Nicht bekannt

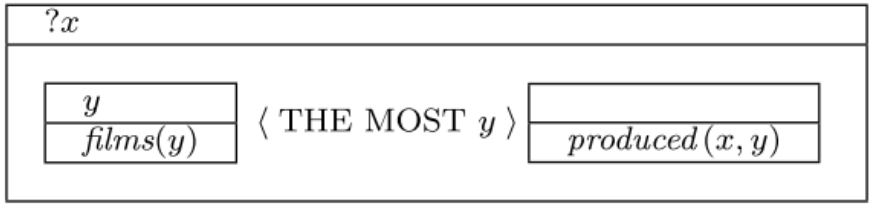
    Können nicht voll spezifiziert werden → On-the-fly
- Benutzung von Stanford POS tagger und Heuristiken
  - Benannte Einträge sind Nominalphrasen → modelliert als Ressource
  - Nomen oft Klasse, aber auch Eigenschaft → daher zwei lexikalische Einträge

    Class-Slot und Property-Slot
  - Verben meistens Eigenschaft → Property-Slot

Syntax und Semantik

- Syntaxbaum
- Semantik





- Semantische Repräsentation → SPARQL Template(s)

Property durch Verb

```
SELECT ?x WHERE {
    ?x ?p ?y .
    ?y rdf:type ?c .
  } ORDER BY DESC(COUNT(?y))
OFFSET 0 LIMIT 1
```

Slots: <?c, class, films>
<?p, property, produced>

Property durch Nomen

```
SELECT ?x WHERE {
    ?x ?p ?y .

  } ORDER BY DESC(COUNT(?y))
OFFSET 0 LIMIT 1
```

Slots: <?p, property, films>

Übersicht

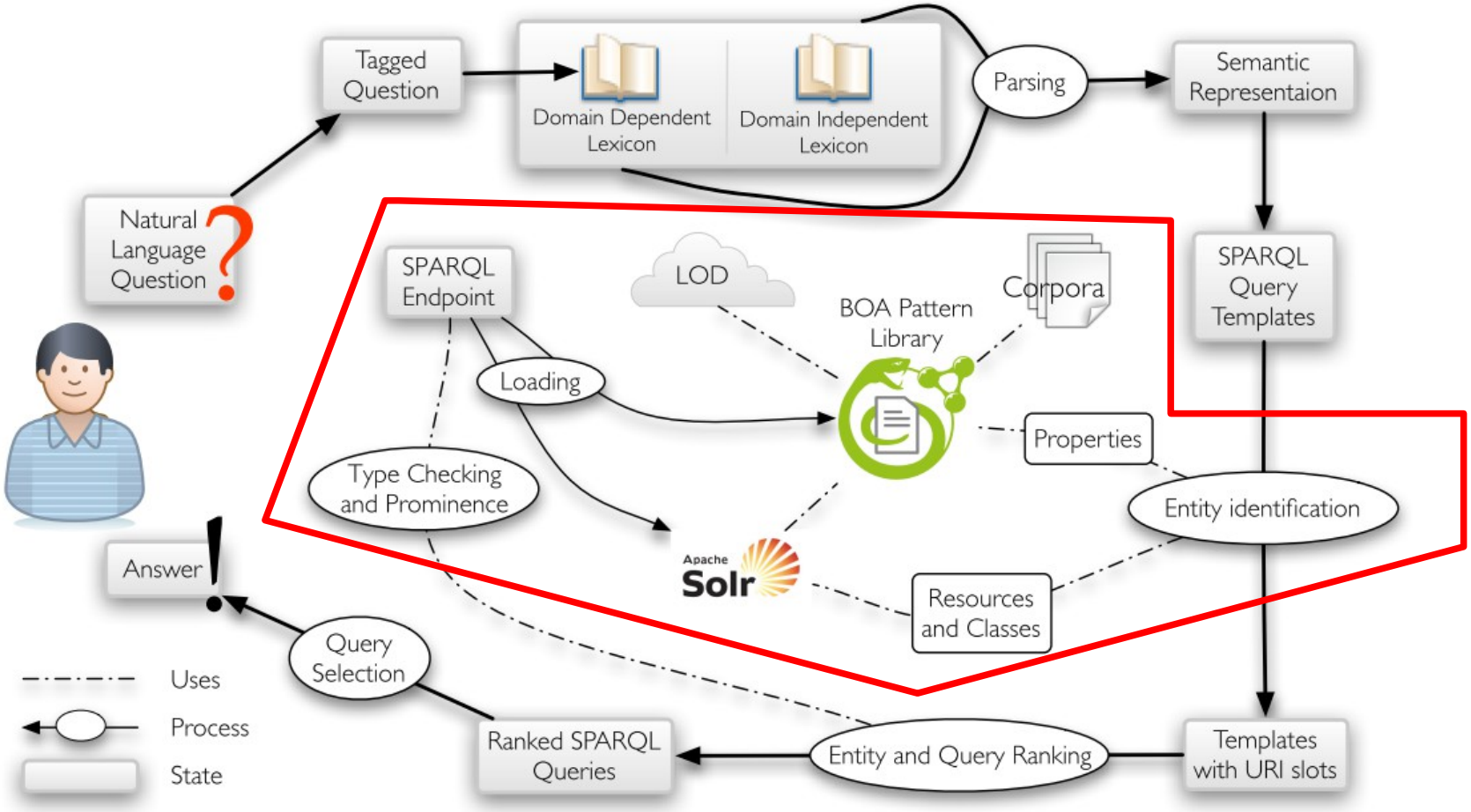


Figure 1: Overview of the template based SPARQL query generator.

Entity Identifikation - Suche

TECHNISCHE UNIVERSITÄT DARMSTADT

- Allgemeiner Ansatz
  - Synonyme aus WordNet → Synsets
  - Filtere Ergebnis
- Property Erkennung

  nur wenn String für ein Property-Label steht
  - „X the creator of Y“ oder

    „Y is a book by X“
  - BOA Pattern Bibliothek
    - Vergleiche String mit NLE

Entity Identifikation - Scoring

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ergebnis sind viele Paare
  - BOA Pattern – Tupel (NLE,Prädikat)
- Beurteilung der Paare
  - Support: Viele BOA Pattern für Property
  - Typicity: stimmt der RDF Typ / die Domäne?
  - Specificity: wenig Pattern Zuordnungen
- Gesamtscore

Entity Identifikation - Scoring

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ergebnis sind viele Paare
  - BOA Pattern – Tupel (NLE,Prädikat)
- Beurteilung der Paare
  - Support: Viele BOA Pattern für Property
  - Typicity: stimmt der RDF Typ / die Domäne?
  - Specificity: wenig Pattern Zuordnungen
- Gesamtscore

$$\mathcal{I}(\mathrm{p}) = \{(x, y) : (x\ \mathrm{p}\ y) \in \mathcal{K}\}$$

„Wir gehen zum Klettern“
„Wir gehen in Django“

Entity Identifikation - Scoring

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ergebnis sind viele Paare
  - BOA Pattern – Tupel (NLE,Prädikat)
- Beurteilung der Paare
  - Support: Viele BOA Pattern für Property
  - Typicity: stimmt der RDF Typ / die Domäne?
  - Specificity: wenig Pattern Zuordnungen
- Gesamtscore

„?D, the creator of ?R“
„?R is a book by ?D“

Entity Identifikation - Scoring

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ergebnis sind viele Paare
  - BOA Pattern – Tupel (NLE,Prädikat)
- Beurteilung der Paare
  - Support: Viele BOA Pattern für Property
  - Typicity: stimmt der RDF Typ / die Domäne?
  - Specificity: wenig Pattern Zuordnungen
- Gesamtscore

„Keramik wird durch sintern hergestellt“

Übersicht

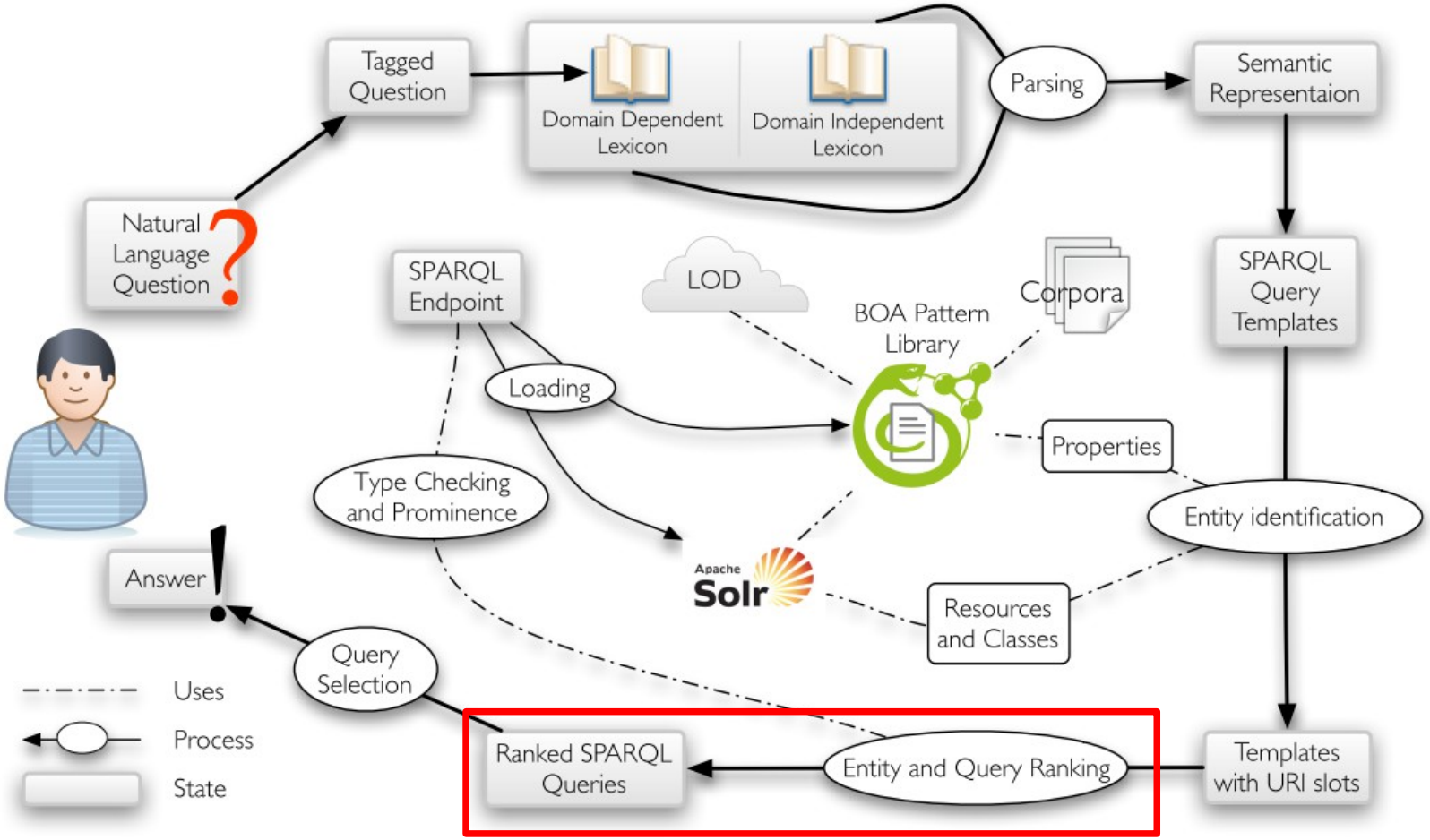


Figure 1: Overview of the template based SPARQL query generator.

Query Ranking...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Entitäten identifiziert
- Einstufen der SPARQL Queries
- Für jeden Slot und ein mögliche Entität:
  - Similarity Score: String ähnlich zum Eingabestring
  - Prominence Score: $$\varphi(e) = \begin{cases} \log_2 |\{(\mathtt{x}, \mathtt{y}) : \mathtt{x}\ \ \mathtt{e}\ \ \mathtt{y}\}| & \text{if } \mathtt{e} \text{ is a property} \\ \log_2 |\{(\mathtt{x}, \mathtt{y}) : \mathtt{x}\ \ \mathtt{y}\ \ \mathtt{e}\}| & \text{else,} \end{cases}$$
- Gesamtscore der Entität mit Gewichtung
- Typ Checks
- Queryscore aus dem Mittel der Kosten

...and Selection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Welche Antwort an den Benutzer geben?
  - Die höchst bewertete Query?
- Test-Query
  - Verwerfen wenn kein Ergebnis
  - Unpassende Entitäten
  - Count-Queries
    - WHERE-Klausel

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- QALD Benchmark
  - $$\text{Recall} = \frac{\text{number of correct resources returned by system}}{\text{number of resources in gold standard answer}}$$

    $$\text{Precision} = \frac{\text{number of correct resources returned by system}}{\text{number of resources returned by system}}$$
  - Manuell korrigierte POS tags
- 50 Fragen
  - 11 Fragen ausgenommen
    - Prädikate nicht eingebunden (YAGO / FOAF)
  - 5 Fragen
    - Unbekannte syntaktische Repräsentation
    - Nicht abgedeckte Domain unabhängiger Ausdrücke

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- 34 verbleibende Fragen
  - 19 exakt beantwortet (P/R = 1.0)
  - 2 fast korrekt (P/R > 0.8)
  - Mittel der Precisionscore 0.61 / der Recallscore 0.63
  - Vergleichbar mit
    - FREyA
    - PowerAqua

Diskussion/Fehleranalyse Templates

TECHNISCHE UNIVERSITÄT DARMSTADT

- Quelle von Fehlern
- Nicht korrekte Templates
  - kein passendes Template
    - „join the EU“ Property Slot: join; Resource: EU

      DBpedia property is: accessioneudate
  - Selbe Problem bei komplexen YAGO Kategorien
  - Vorverarbeitungsschritt

Diskussion/Fehleranalyse Entitätenidentifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Property oder Ressource class nicht gefunden
  - Keine Repräsentation in der NL
    - „Give me all movies *with* Tom Cruise.“

      Annahme: Domain „Movie“

      „With“ → „onto:starring“
  - Nicht Teil des Ansatzes
- Repräsentation in der NL
  - Beispiel: „Inhabitants“, „owns“ und „higher

    „populationTotal“, „keyPerson“ und „elevationM“
  - Benötigen tiefer greifende Semantische Analyse

Diskussion/Fehleranalyse
Query Auswahl

TECHNISCHE UNIVERSITÄT DARMSTADT

- Falsche Entität, obwohl die richtige dabei
  - „Who wrote The pillars of the Earth?“
    - onto:writer (matched) → onto:author
    - Deswegen

      „Res:The_Pillars_of_the_Earth(TV-Miniseries)“
- Zu wenig Informationen
  - „Founded“ hat viele Properties

    foundingYear, foundationPerson, foundationPlace...
  - Hinweise erkennen
    - Enthält Zahl
    - Enthält „Wh-Wort“

Diskussion/Fehleranalyse
Andere Gründe
Prototype

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zu wenig Ergebnisse
  - „What languages are spoken in Estonia?“
  - Richtiges Property
    - 1.0 Precision, geringer Recall
    - Gold-Query verwendet UNION Operator
  - Erweiterte Suche

- Prototype
  - Knowledge Base: DBpedia
  - Mit Benutzerfeedback (passende alternative property values)
  - Offline preprocessed Indizes
    - Example at: http://autosparql-tbsl.dl-learner.org/

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- PowerAqua
  - Keine Annahmen über Vokabular oder Datenstruktur
  - Focus auf Skalierbarkeit
  - Stark bei großen heterogenen Datensets
  - Probleme bei YAGO Kategorien
  - Keine tiefe linguistische Analyse („more than“, “the most“,...)
- Pythia
  - Tiefe linguistische Analyse
  - Kann mit „more than“ usw. umgehen
  - Basiert auf manuellem Lexikon

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- FREyA
  - Entitätenauswahl mit Benutzerunterstützung
  - Benutzer kennt sich nicht mit Modellierung und Vokabular aus
- Andere Keyword basierende Ansätze
  - Viele Systeme in den letzten Jahren
  - Semantische Suchmaschinen: Swoogle, Watson, Sigma and Sindice
  - Indiziert RDF Daten und Subgraphen
  - Vorgefertigte SPARQL Queries

Zusammenfassung und weiteres Vorgehen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Neuer Ansatz
  - Basiert auf tiefer Linguistischer Suche
  - SPARQL Templates with Slots
  - Entitäten durch Stringähnlichkeiten und NLP
  - Repräsentiert NL Eingabe
    - Quantifizierer und Superlative
- Weiteres Vorgehen
  - Starre Templates → flexiblere Verarbeitung
  - Preprocessing Schritt für komplexe Kategorien
  - Fallback Strategie
    - Mittels Active Learning bei inkorrekten / unvollständigen Templates
    - Benutzerfeedback

Vision
Meine Meinung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Testen des Ansatzes auf verschiedenen Domänen
  - DBpedia 100 Trainingsdaten / 100 Testdaten
  - MusicBrainz
- Kleine Benutzbarkeitsstudie
- Ultimatives Ziel
  - heterogene Wissensdatenbank

- IMHO
  - Interessanter, vielversprechender Ansatz
  - (Teilweise) gute Ergebnisse
  - Benutzer Feedback
  - Unterschiedlichen Sprachen

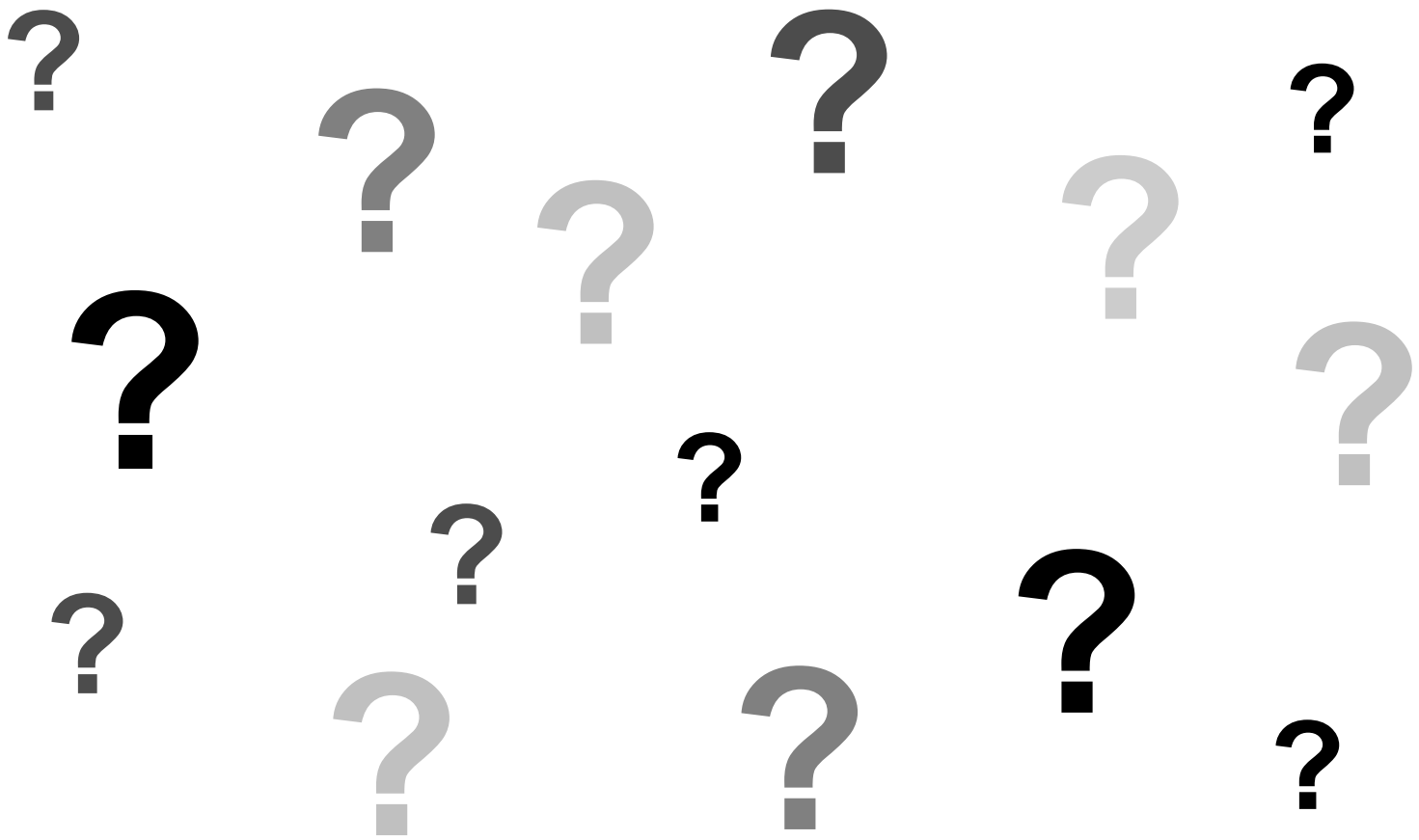
Fragen

TECHNISCHE UNIVERSITÄT DARMSTADT

id	question	precision	recall
2	Who has been the 5th president of the United States of America		
4	Who was Tom Hanks married to	1.0	1.0
5	Which people were born in Heraklion	0.91	1.0
7	Which companies work in the aerospace industry as well as on nuclear reactor technology		
8	Which people have as their given name Jimmy		
9	Who developed the video game World of Warcraft	1.0	1.0
10	Who was the wife of president Lincoln	1.0	1.0
12	Which caves have more than 3 entrances	1.0	1.0
13	Which cities have more than 2000000 inhabitants	0.04	0.26
14	Who owns Aldi		
16	Give me all soccer clubs in the Premier League	0.5	0.86
17	In which programming language is GIMP written	1.0	1.0
18	What languages are spoken in Estonia	1.0	0.14
20	Which country does the Airedale Terrier come from	1.0	1.0
21	What is the highest mountain	1.0	1.0
24	Which organizations were founded in 1950	0.0	0.0
25	Which genre does DBpedia belong to	1.0	1.0
26	When was DBpedia released	1.0	1.0
27	Who created English Wikipedia	1.0	1.0
28	Which companies are located in California USA	0.8	0.76
30	How many films did Leonardo DiCaprio star in	1.0	1.0
31	Who produced the most films	1.0	1.0
32	Is Christian Bale starring in Batman Begins	1.0	1.0
33	Which music albums contain the song Last Christmas		
34	Give me all films produced by Hal Roach	1.0	1.0
35	Give me all actors starring in Batman Begins	1.0	0.86
36	Give me all movies with Tom Cruise	0.08	0.75
37	List all episodes of the first season of the HBO television series The Sopranos		
38	books by Da...		

Fragen

TECHNISCHE UNIVERSITÄT DARMSTADT



Quellen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Paper http://www2012.wwwconference.org/proceedings/proceedings/p639.pdf
- RDF http://www.w3.org/RDF/
  SPARQL http://www.w3.org/TR/rdf-sparql-query/
  SPARQL http://www.lukeslog.de/?page_id=841
  POS tagger http://nlp.stanford.edu/software/tagger.shtml
  synset http://wordnet.princeton.edu/
  Prototype http://autosparql-tbsl.dl-learner.org/
- Pictures http://www.alchemyapi.com/api/entity/ldata.html