A Tutorial on Question Answering Systems

Saeedeh Shekarpour

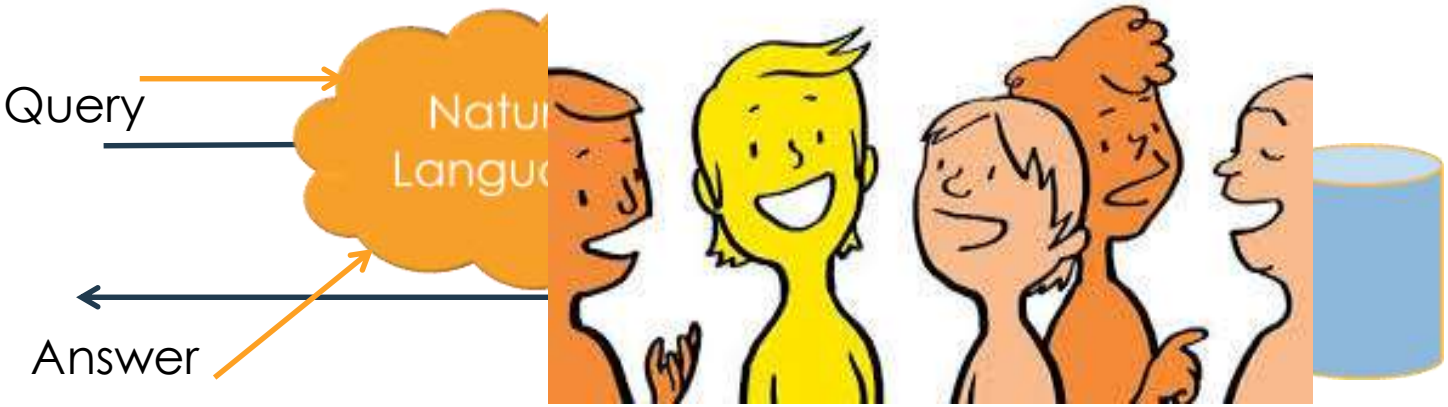
Postdoc researcher from EIS research group

Outline

- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

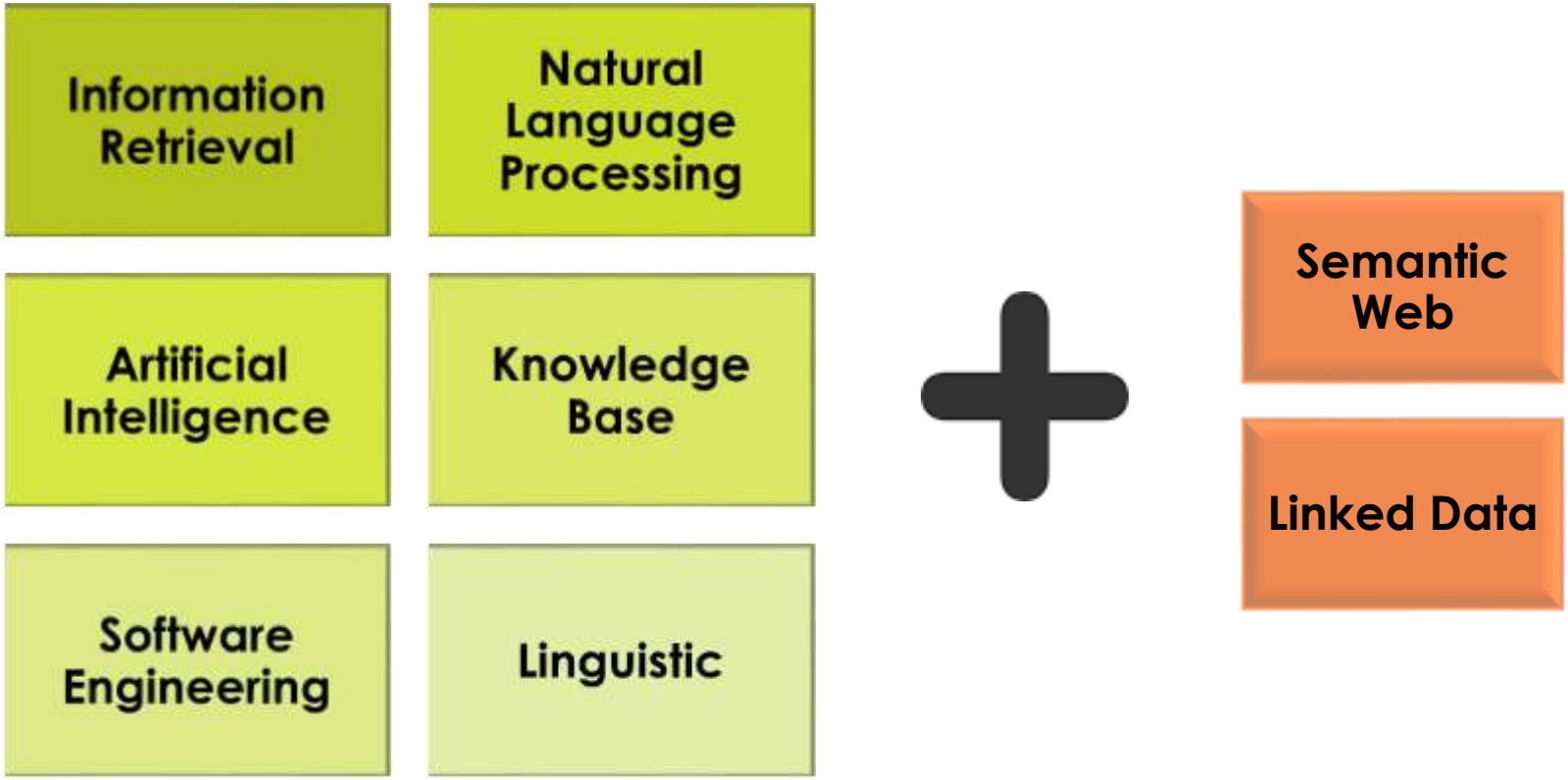
What is a Question Answering (QA) system?

Systems that automatically answer questions posed by humans in natural language query.



Natural language is the common way for sharing knowledge

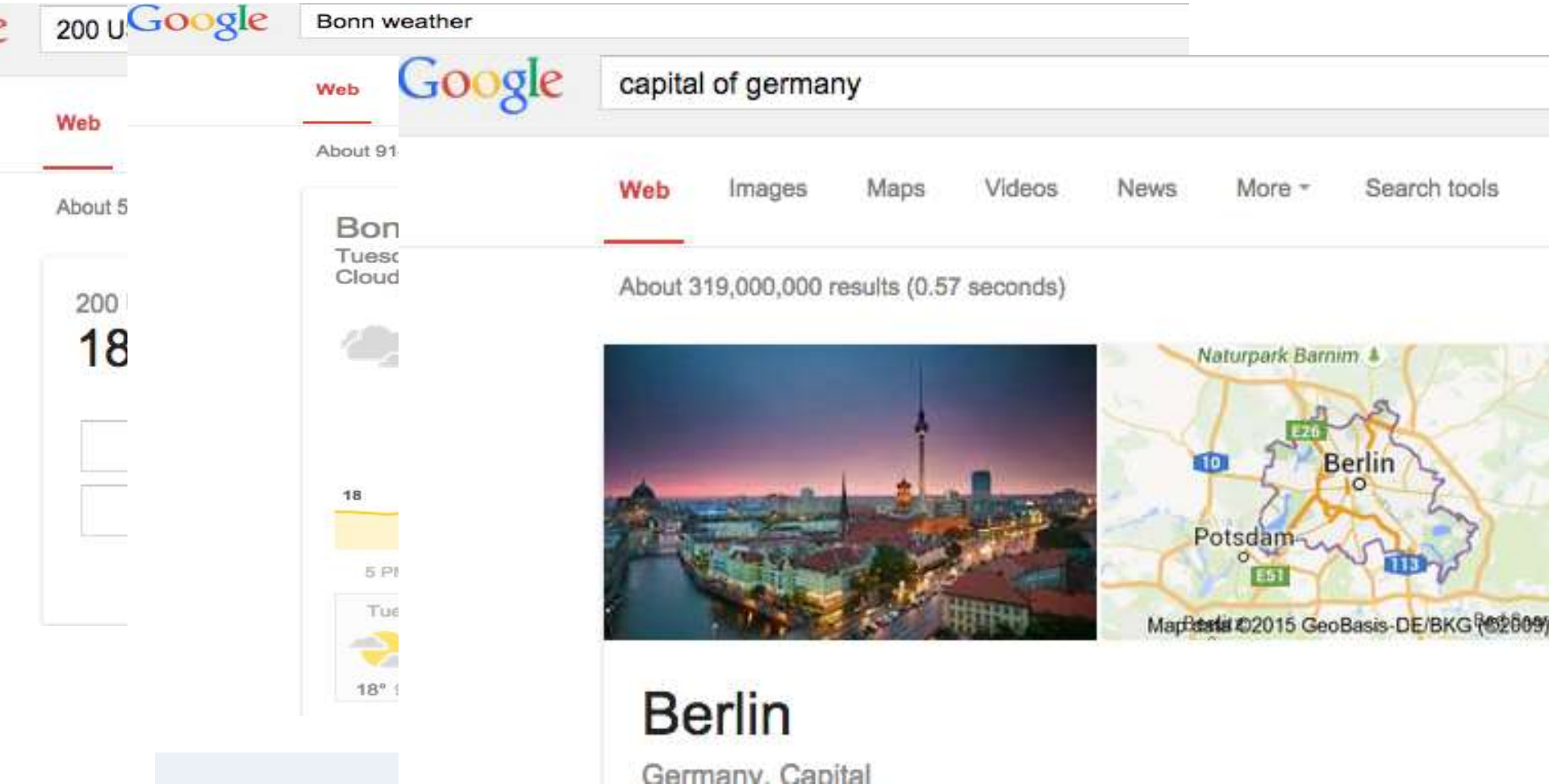
Question answering is a multi disciplinary field



Difference to Information Retrieval

- Information Retrieval is a **query driven** approach for accessing information.
  - System returns a list of documents.
  - It is responsibility of user to navigate on the retrieved documents and find its own information need.
- Question Answering is an **answer driven** approach for accessing information.
  - User asks its question in natural language (i.e. phrase-based, full sentence or even keyword based) queries.
  - System returns the list of short answers.
  - More complex functionality.

Search engines are moving towards QA



Search engines still lack the ability to answer more complex queries

Google

researchers in breast cancer

Web News Images Videos Shopping More Search tools

About 44,800,000 results (0.64 seconds)

Breast Cancer Research - ClimbForHope.com
Ad www.climbforhope.com/
Aid Revolutionary **Cancer Research**. Make the Difference. Save Lives.

Scholarly articles for **researchers in breast cancer**
Reproductive factors and **breast cancer**. - Kelsey - Cited by 1188
Oral contraceptives in relation to **breast cancer**. - Malone - Cited by 122
… biopsy to avoid axillary dissection in **breast cancer** with … - Veronesi - Cited by 1956

Breast Cancer Research
breast-cancer-research.com/
Online and print journal covering topics of basic and clinical **research** relevant to **breast cancer**. **Research** articles are free to all users.

Natural language queries are classified into different categories

- **Factoid queries**: WH questions like when, who, where.
- **Yes/ No queries**: Is Berlin capital of Germany?
- **Definition queries**: what is leukemia?
- **Cause/consequence queries**: How, Why, What. what are the consequences of the Iraq war ?

Natural language queries are classified into different categories

- **Procedural queries**: which are the steps for getting a Master degree?
- **Comparative queries**: what are the differences between the model A and B?
- **Queries with examples**: list of hard disks similar to hard disk X.
- **Queries about opinion**: What is the opinion of the majority of Americans about the Iraq war?

Corpus Type

- Structured data (relational data bases, RDF knowledge bases).
- Semi-structured data (XML databases)
- Free text
- Multimodal data: image, voice, video

Types of QA systems

- Open-domain: domain independent QA systems can answer any query from any corpus
  - \+ covers wide range of queries
  - \- low accuracy
- Closed-domain: domain specific QA systems are limited to specific domains
  - \+ High accuracy
  - \- limited coverage over the possible queries
  - \- Needs domain expert

Is QA system a need for user?

Search engines query log analysis shows that

Type of query	Query log analysis
Informational	48%
Navigational	20%
Transactional	30%

Real informational queries in Google:

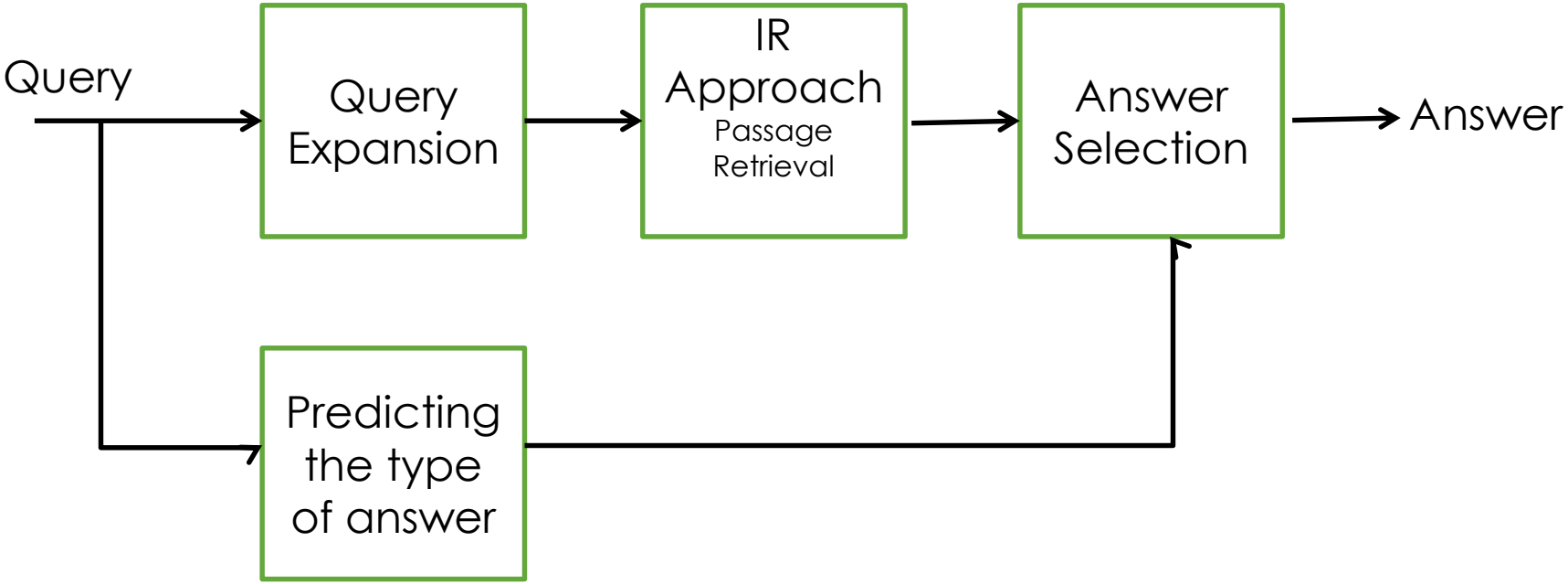
- Who first invented rock and roll music?
- When was the mobile phone invented?
- Where was the hamburger invented?
- How to lose weight?

- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

Text Retrieval Conference (Trec)

- In 1999, Trec initiated a QA track,
- From 1999-2002, participant systems were allowed to return ranked answer snippets. These snippets of text contain the actual answer.
- From 2002, participant systems were allowed to return only the exact answer with confidence rate.
- The evaluation metric was mean reciprocal rank (MRR).

Typical pipeline for the participating systems in Trec



QA at Clef

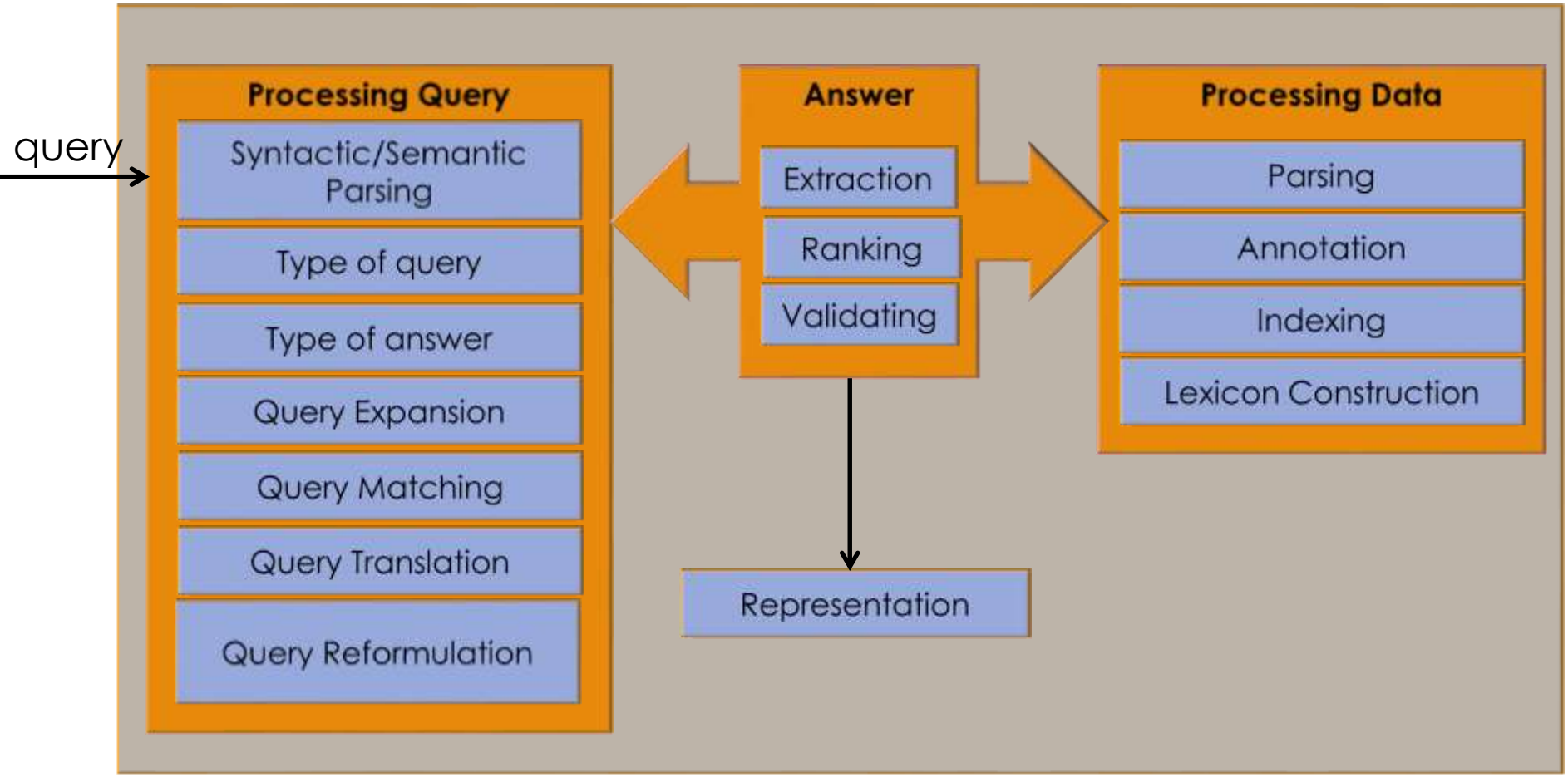
The Cross-Language Evaluation Forum (CLEF) initiative provides

- Since 2003, Clef included a QA track.
- an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts.

Outline

- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

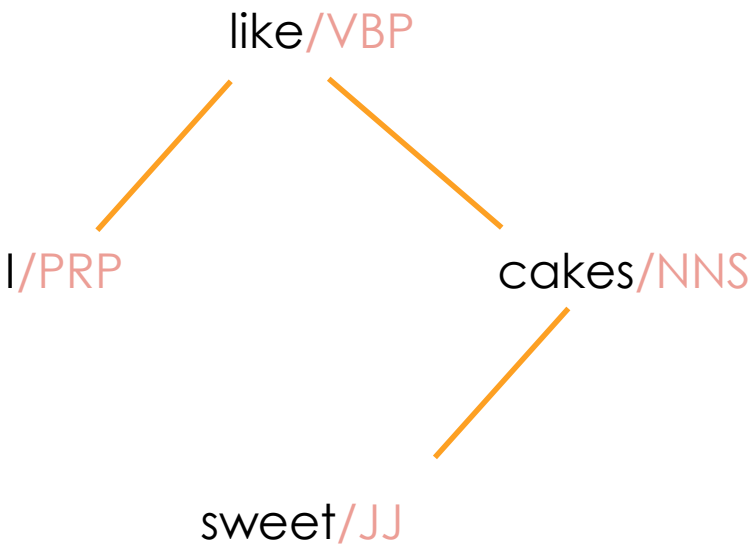
Core of a QA system



Syntactic Parsing: Part-of-speech Tagging

I like sweet cakes

I/PRP like/VBP sweet/JJ cakes/NNS



Type of Answer

Where was the hamburger invented?

Place

White Castle traces the origin of the hamburger to Hamburg, Germany with its invention by Otto Kuase.

Named Entity Recognition on Query

where was Franklin Roosevelt born?

Named Entity: Person

Relation Extraction

Barack Hussein Obama is the 44th and current President of the United States. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School.

Named Entity: Place

Named Entity: Person

Relation: President of

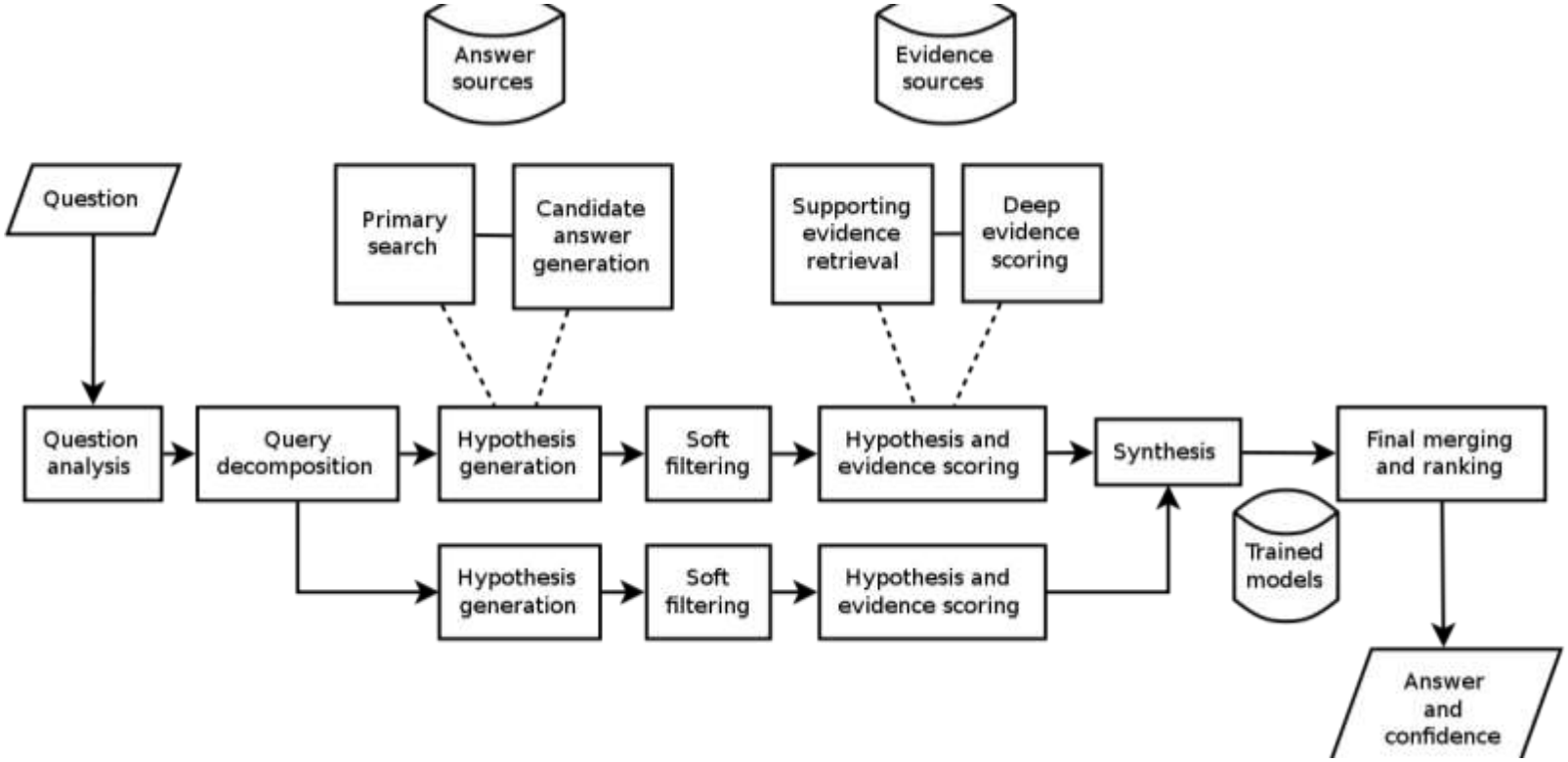
Watson Project

- Watson is a computer which is capable of answering question issued in natural language.
- Questions come from quiz show called Jeopardy.
- The software of this project is called DeepQA project.
- In 2011, Watson won the former winners of quiz show Jeopardy.

Watson Description

- **Hardware:** Watson system has 2,880 POWER processor threads and has 16 terabytes of RAM.
- **Data**: encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies such as DBPedia, WordNet, and Yago were used.
- **Software**: DeepQA software and the Apache UIMA framework. The system was written in various languages, including Java, C++, and Prolog, and runs on the SUSE Linux Enterprise Server 11 operating system using Apache Hadoop framework to provide distributed computing.

The high level architecture of DeepQA



Decomposition example

Query: Of the four countries in the world that the United States does not have diplomatic relations with, the one that's farthest north.

Bhutan, Cuba, Iran, North Korea

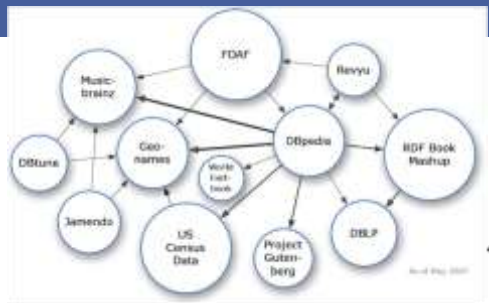
North Korea

Outline

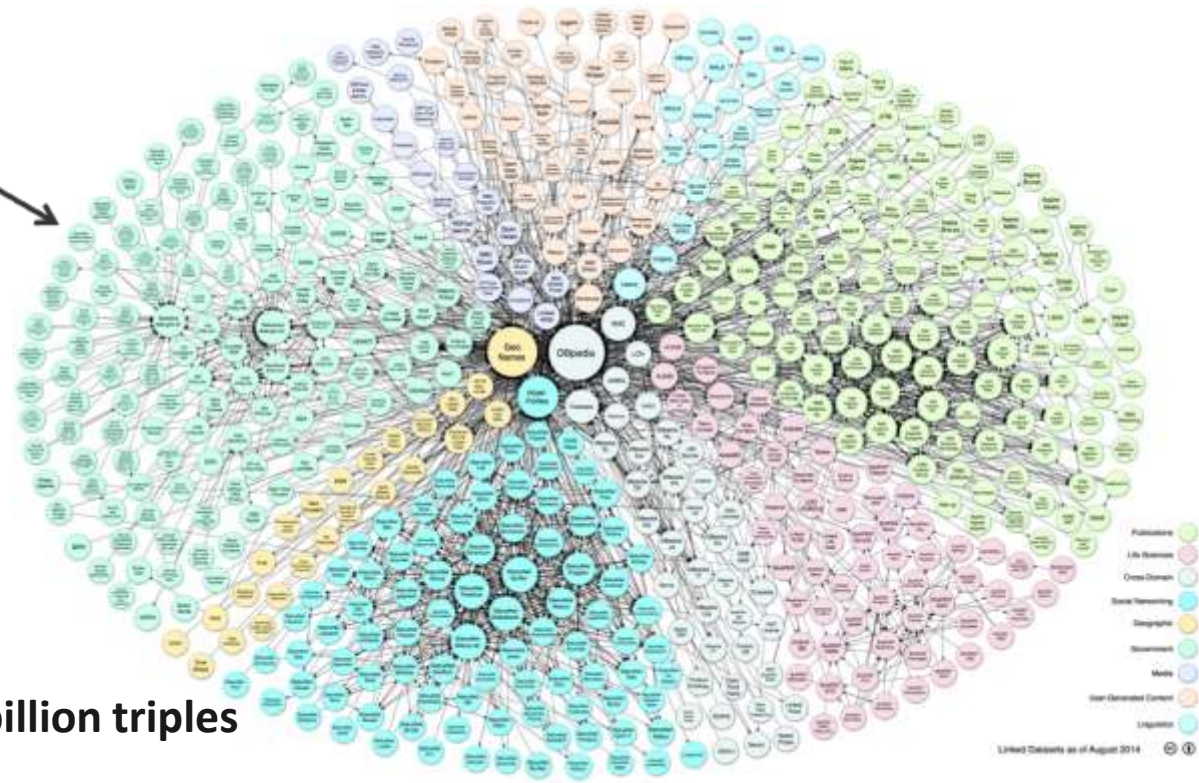
- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

Evolution of Web

The growth of Linked Open Data



May 2007
12 Datasets



August 2014
570 Datasets
More than 74 billion triples

How to retrieve data from Linked Data?

Linked Data characteristics:

- Wide range of topical domains
- Variety in vocabularies
- Interlinked data

SPARQL queries:

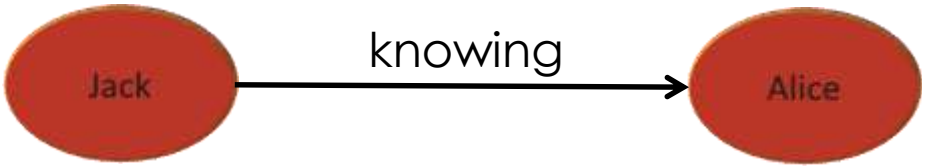
- Knowledge about the ontology
- Proficiency in formulating formal queries
- Explicit and unambigious semantics

Text queries (either keyword or natural language):

- Simple retrieval approach
- Implicit and ambiguous semantics
- Popular

RDF Model

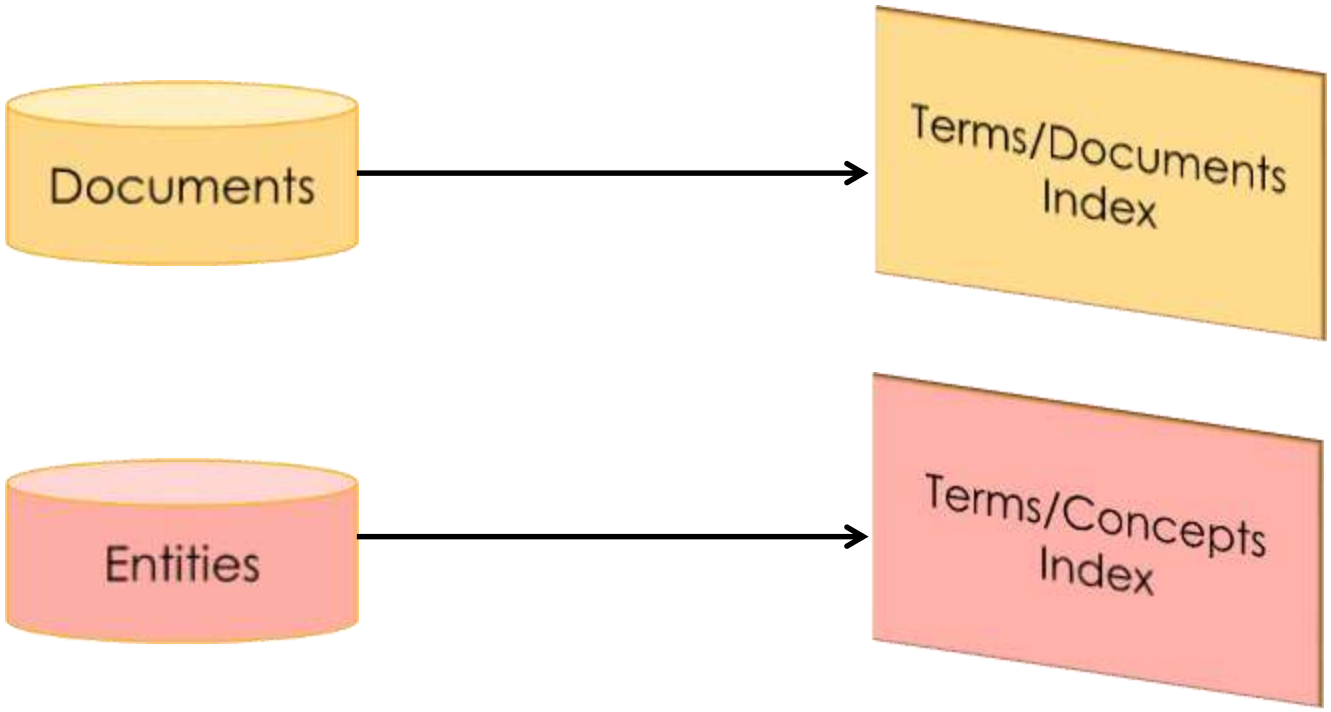
- RDF is an standard for describing Web resources.
- The RDF data model expresses statements about Web resources in the form of subject-predicate-object (triple).
- The statement "Jack knows Alice" is represented as:



Outline

- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

Semantic Indexing



Semantic Annotation

- Name Entity Recognition

Where is the capital of Germany?

Named Entity: Place

- Semantic Annotation

Where is the capital of Germany?

Entity: http://dbpedia.org/resource/Germany

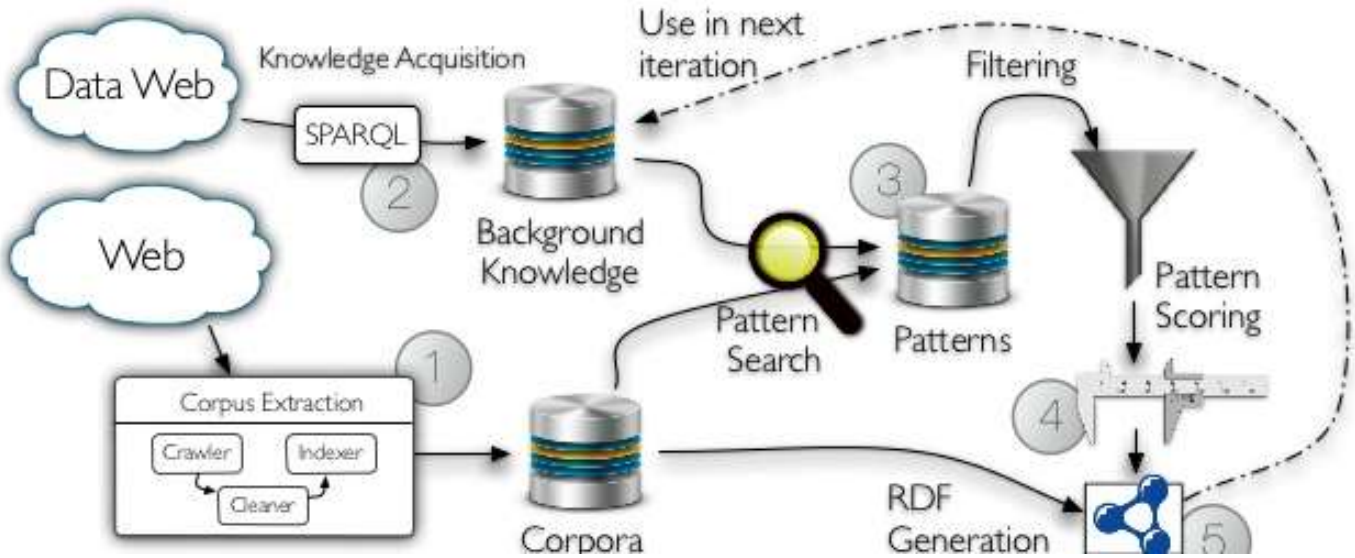
DBpedia Spotlight

Berlin in the 1920s was the third largest municipality in the world. After World War II, the city became divided into East Berlin -- the capital of East Germany -- and West Berlin, a West German exclave surrounded by the Berlin Wall from 1961–89.

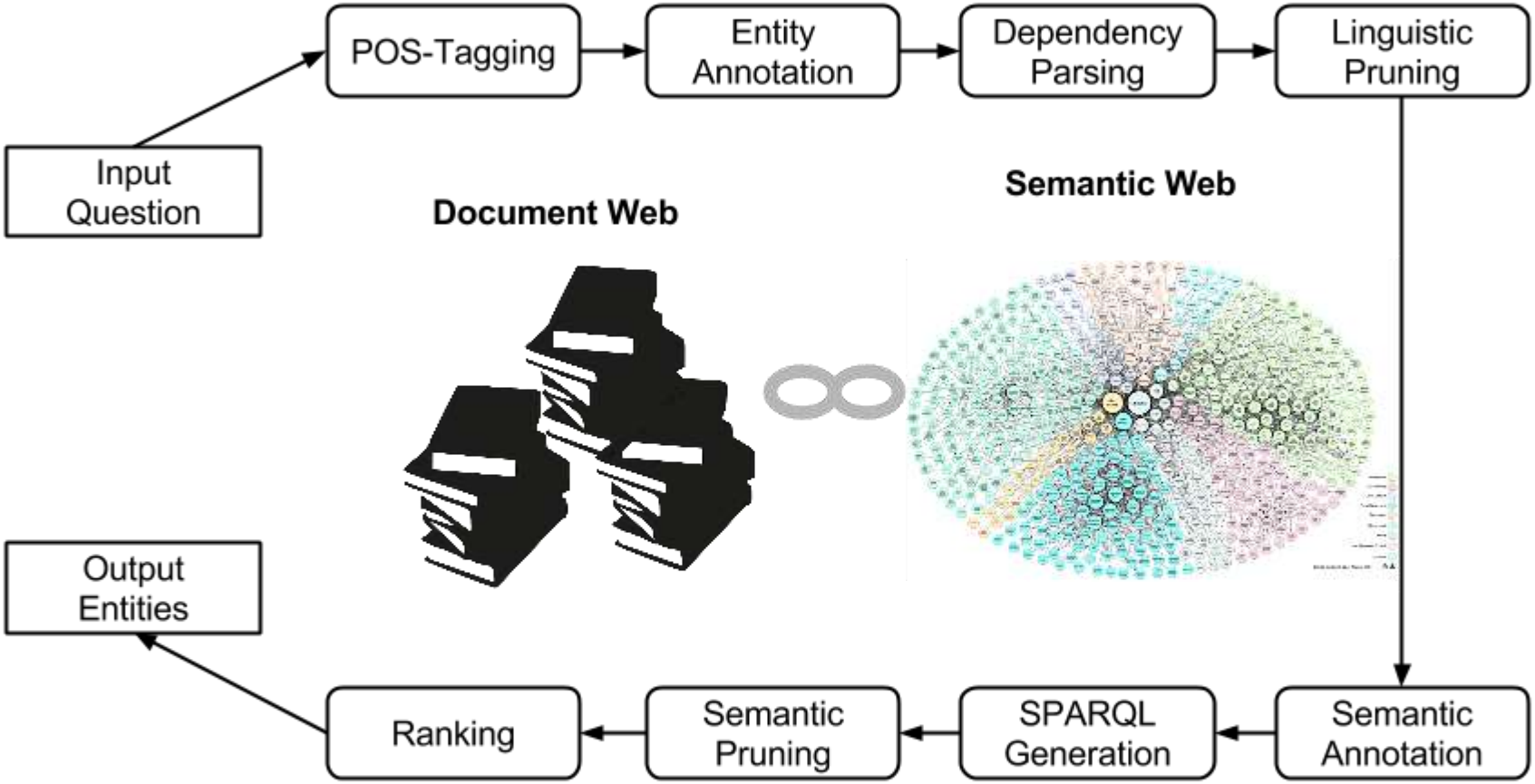
Berlin in the 1920s was the third largest municipality in the world. After World War II, the city became divided into East Berlin -- the capital of East Germany -- and West Berlin, a West German exclave surrounded by the Berlin Wall from 1961–89.

Relation extraction leveraging Data Web: BOA Library

The BOA approach



HAWK: Hybrid Question Answering over Linked Data



TBSL: Template-based SPARQL generator

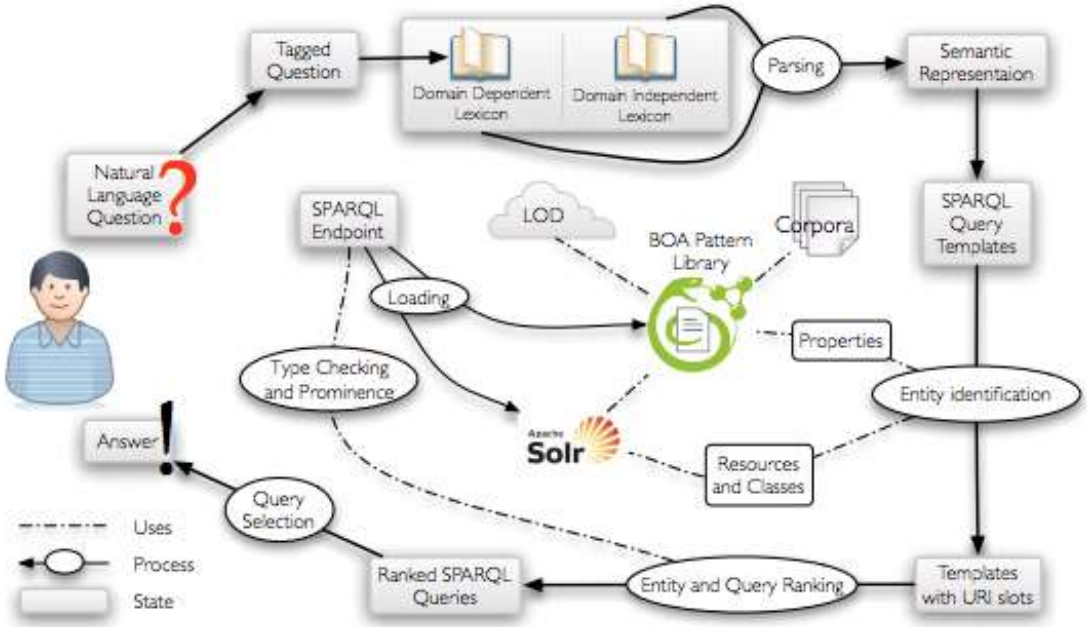
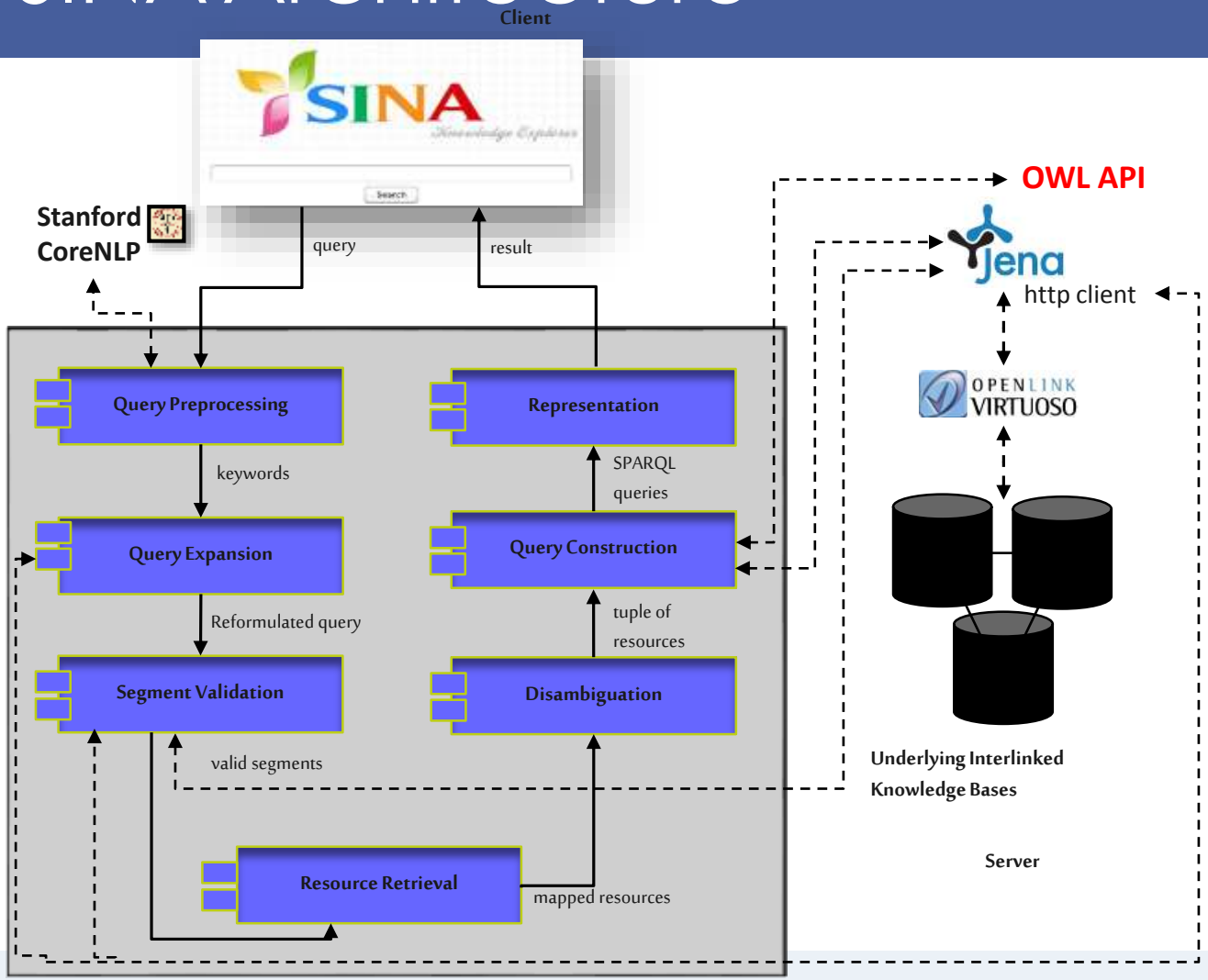


Figure 1: Overview of the template based SPARQL query generator.

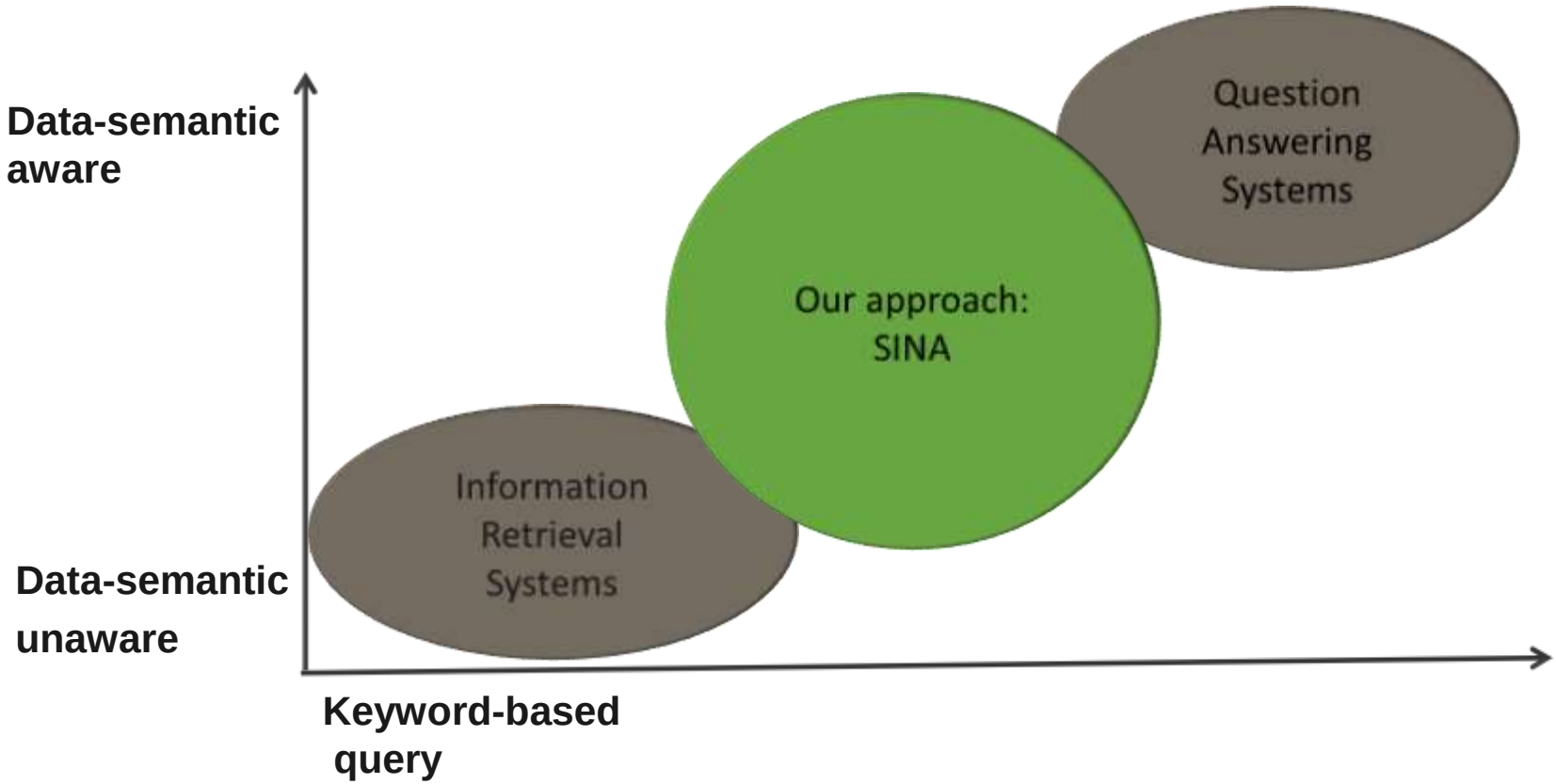
Outline

- Introduction
- Associations for evaluation QA systems
- Preliminary Concepts
- Data Web
- Emerging Concepts
- Deeper view on SINA Project

SINA Architecture



Comparison of search approaches



Objective: transformation from textual query to formal query

1 Which televisions shows were created by Walt Disney?

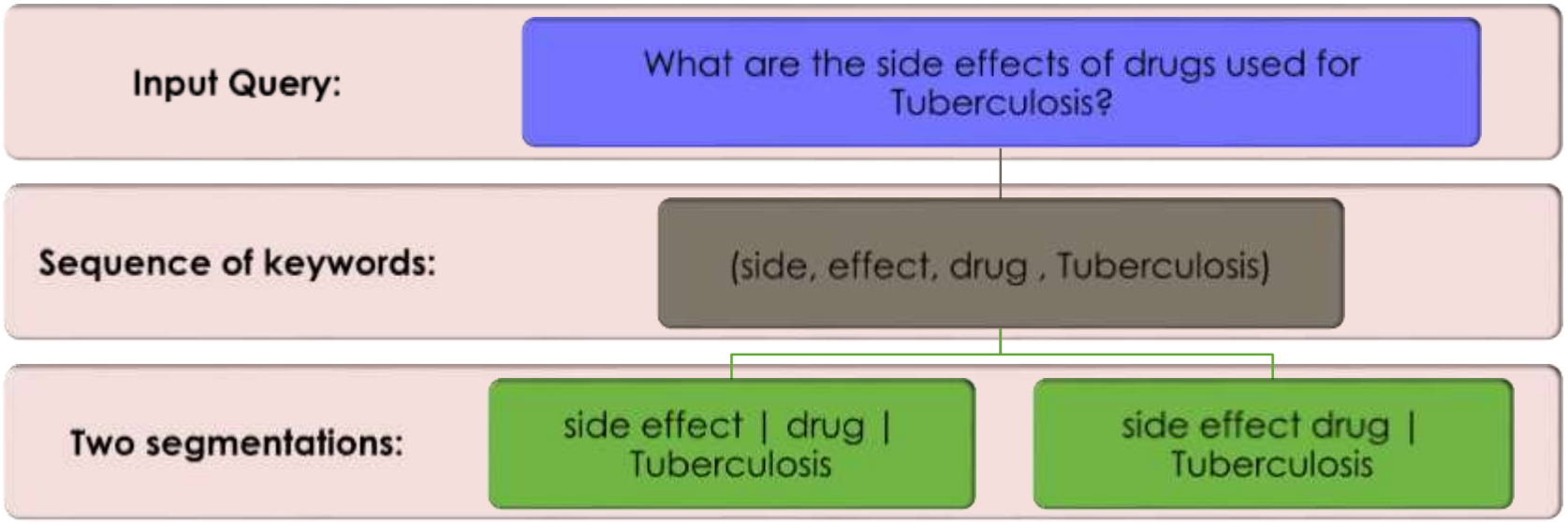
2
```
SELECT * WHERE
{ ?v0 a              dbo:TelevisionShow.
  ?v0 dbo:creator    dbr:Walt_Disney. }
```

3


The addressed challenges in SINA

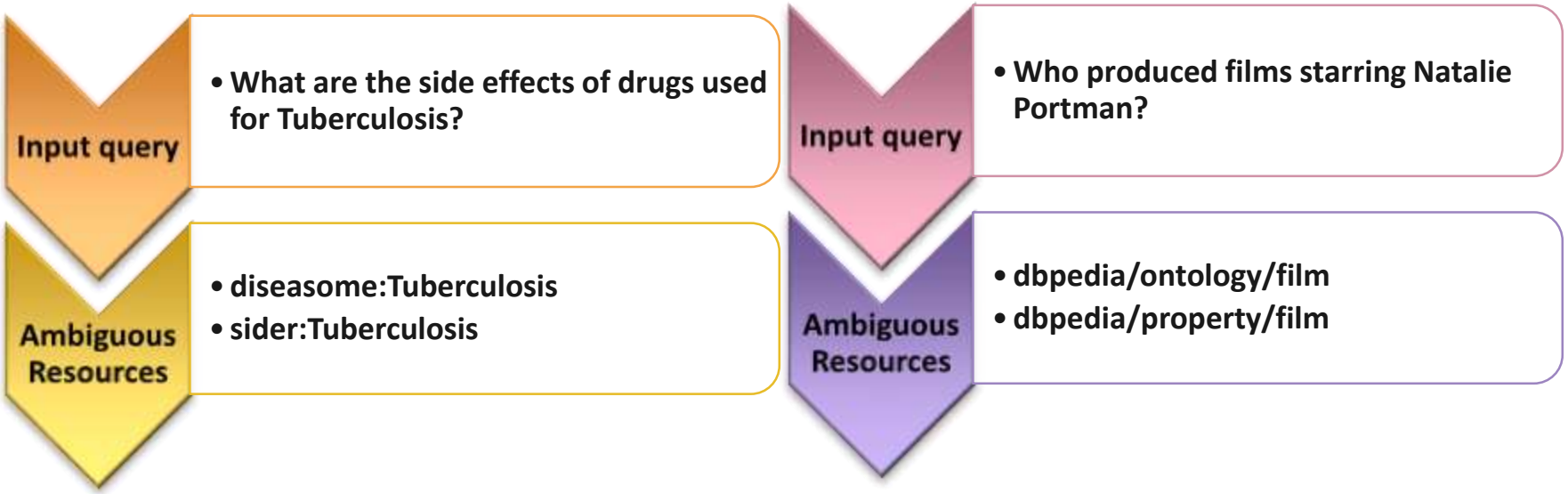
Query Segmentation

- **Definition:** query segmentation is the process of identifying the right segments of data items that occur in the keyword queries.

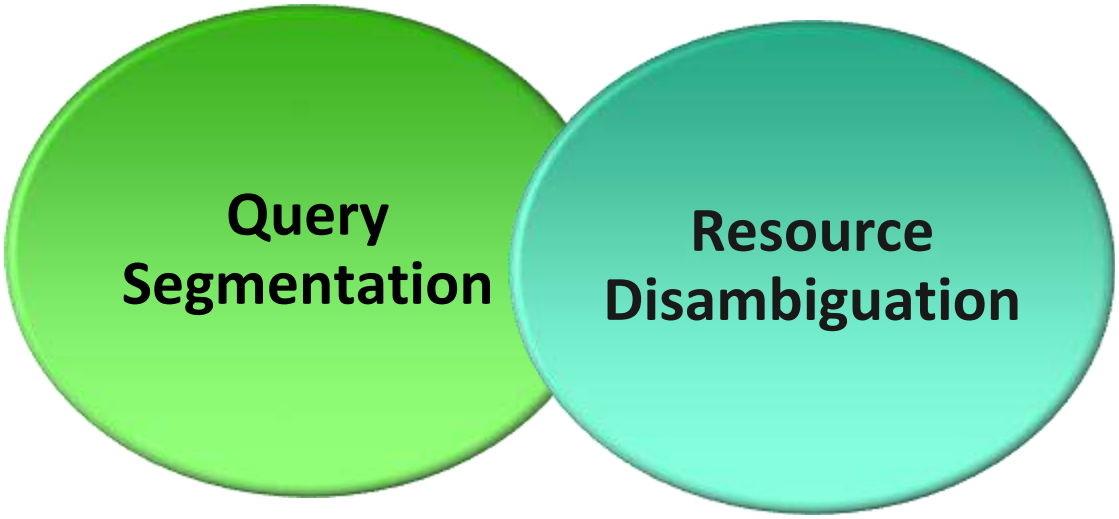


Resource Disambiguation

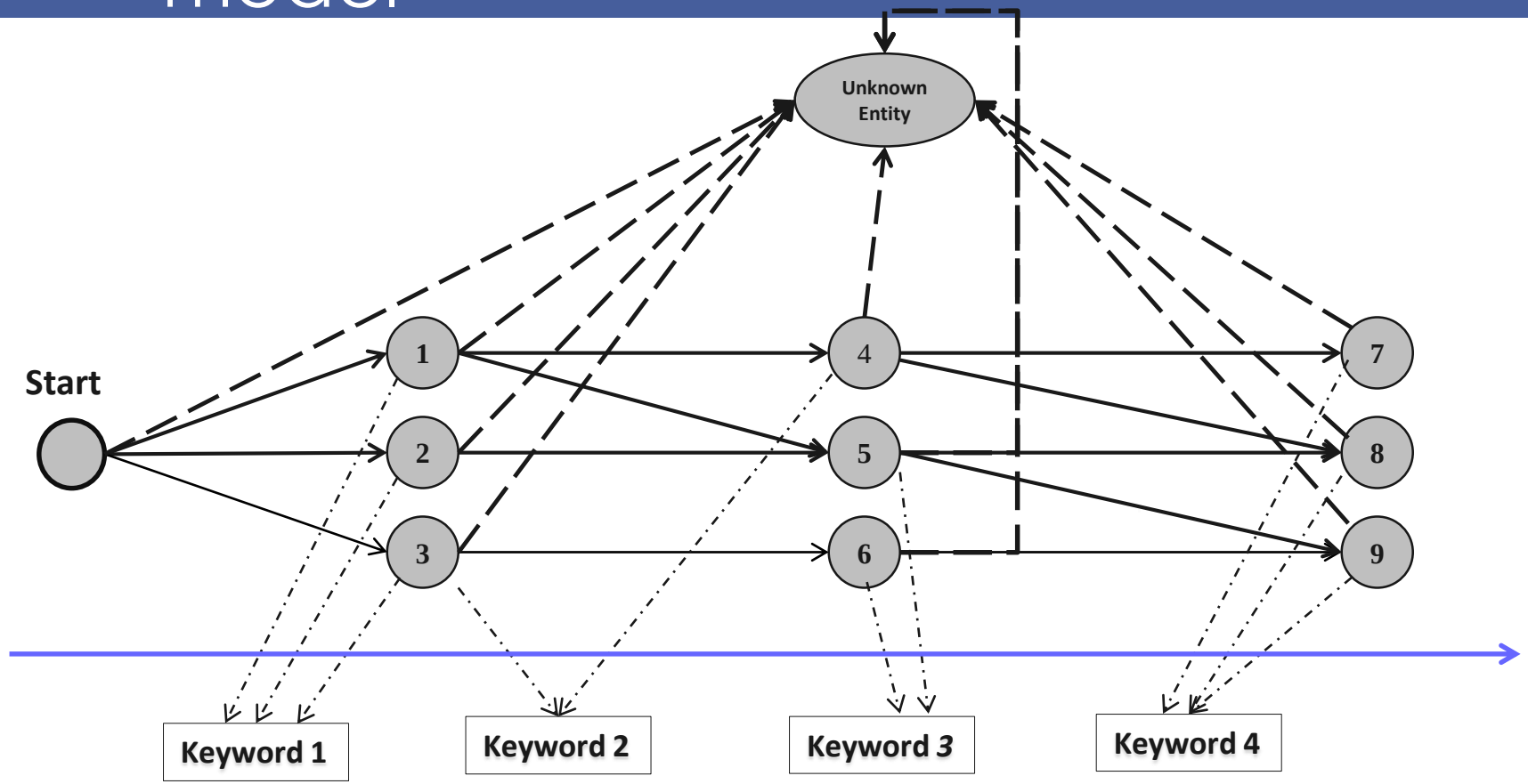
- **Definition**: resource disambiguation is the process of recognizing the suitable resources in the underlying knowledge base.



Concurrent Approach



Modeling using hidden Markov model



Bootstrapping the model parameters

1. Emission probability is defined based on the similarity of the label of each state with a segment, this similarity is computed based on string-similarity and Jaccard-similarity.

2. Semantic relatedness is a base for transition probability and initial probability. Intuitively, it is based on two values: distance and connectivity degree. We transform these two values to hub and authority values using weighted HITS algorithm.

3. HITS algorithm is a link analysis algorithm that was originally developed for ranking Web pages. It assign a hub and authority value to each web page.

4. Initial probability and transition probability are defined as a uniform distribution over the hub and and authority values.

Output of the model

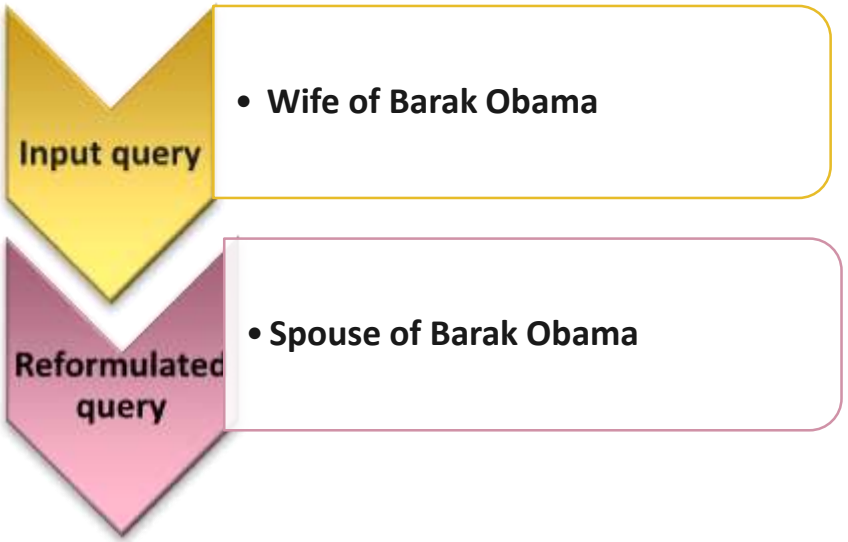
Sequence of keywords (television show creat Walt Disney)

Paths

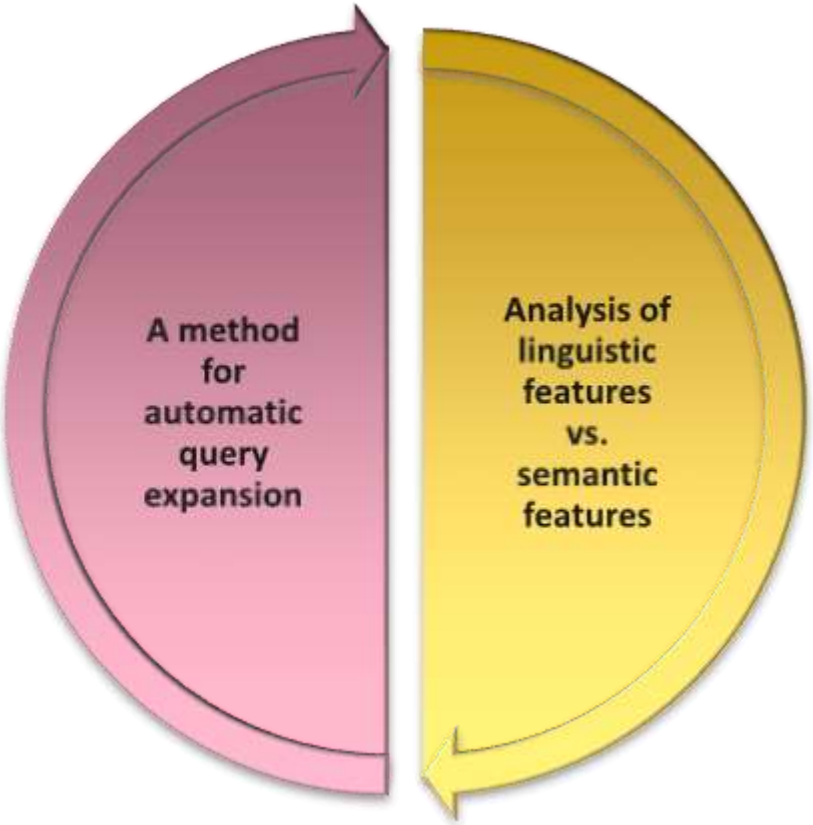
0.0023	dbo:TelevisionShow		dbo:creator	dbr:Walt_Disney
0.0014	dbo:TelevisionShow		dbo:creator	dbr:Category:Walt_Disney
0.000589	dbr:TelevisionShow		dbo:creator	dbr:Walt_Disney
0.000353	dbr:TelevisionShow		dbo:creator	dbr:Category:Walt_Disney
0.0000376	dbp:television	dbp:show	dbo:creator	dbr:Category:Walt_Disney

Query Expansion

- **Definition:** query expansion is a way of reformulating the input query in order to overcome the vocabulary mismatch problem.



Query Expansion



Linguistic features

- **WordNet** is a popular data source for expansion.
- Linguistic features extracted from **WordNet** are:
  1. **Synonyms:** words having a similar meanings to the input keyword.
  2. **Hyponyms:** words representing a specialization of the input keyword.
  3. **Hypernyms:** words representing a generalization of the input keyword.

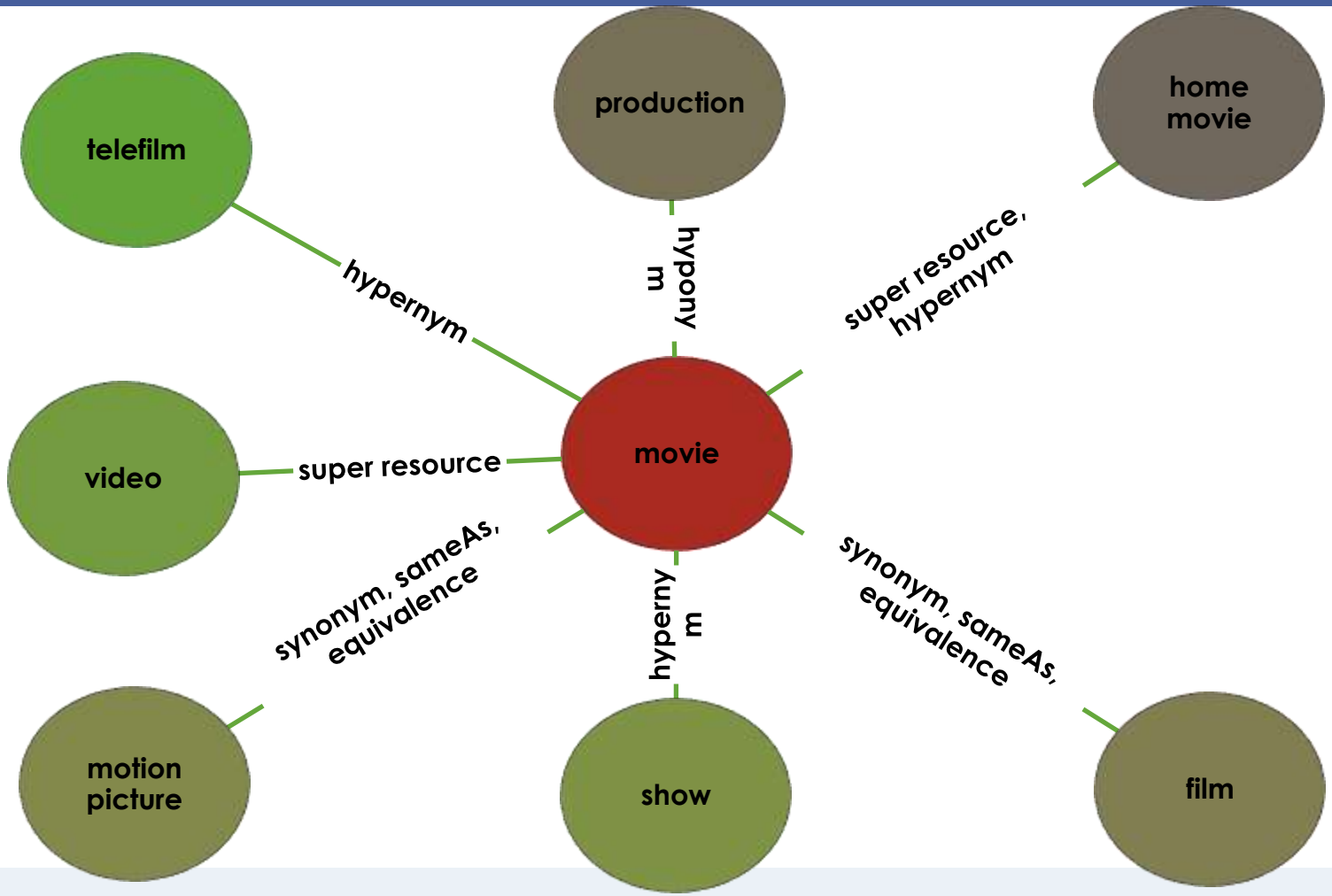
Semantic features from Linked Data

1. **SameAs**: deriving resources using owl:sameAs.
2. **SeeAlso**: deriving resources using rdfs:seeAlso.
3. **Equivalence class/property:** deriving classes or properties using owl:equivalentClass and owl:equivalentProperty.
4. **Super class/property**: deriving all super classes/properties of by following the rdfs:subClassOf or rdfs:subPropertyOf property.
5. **Sub class/property**: deriving resources by following the rdfs:subClassOf or rdfs:subPropertyOf property paths ending with the input resource.

Semantic features from Linked Data

5. **Broader concepts**: deriving using the SKOS vocabulary properties skos:broader and skos:broadMatch.
6. **Narrower concepts:** deriving concepts using skos:narrower and skos:narrowMatch.
7. **Related concepts**: deriving concepts using skos:closeMatch, skos:mappingRelation and skos:exactMatch.

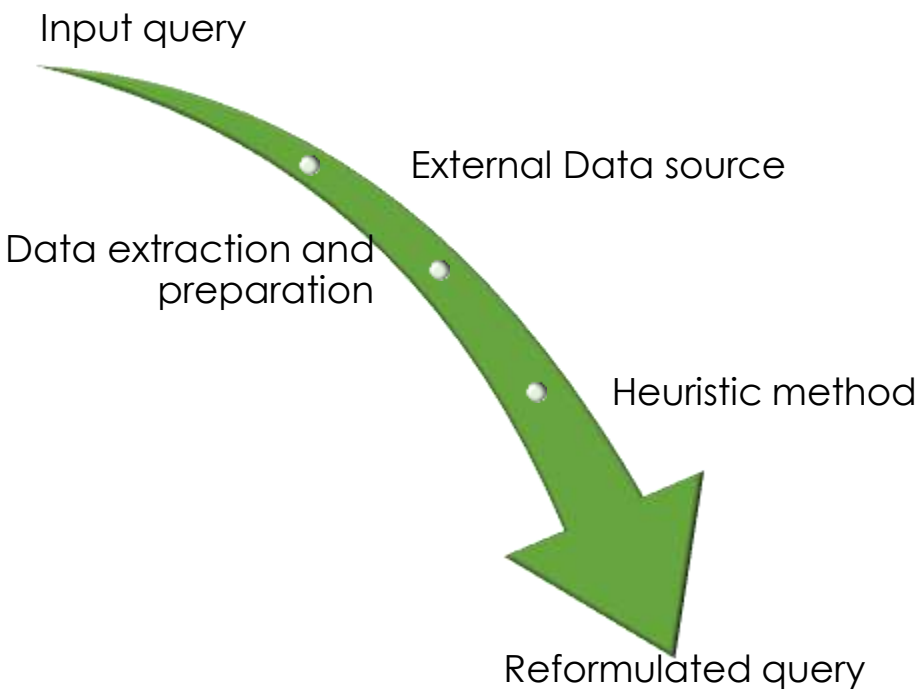
Exemplary expansion graph of the word movie



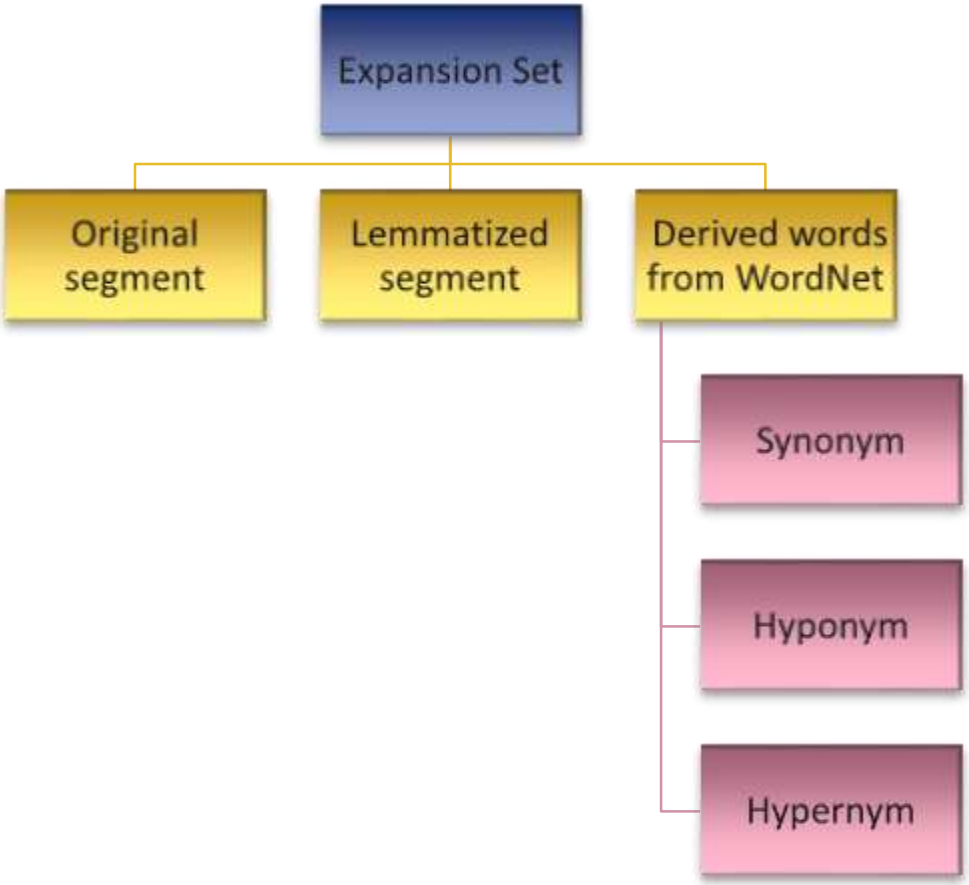
Statistics over the number of the derived words from WordNet and Linked Data

Feature	#derived words
synonym	503
hyponym	2703
hypernym	657
sameAs	2332
seeAlso	49
equivalence	2
super class/property	267
Sub class/property	2166

Automatic query expansion

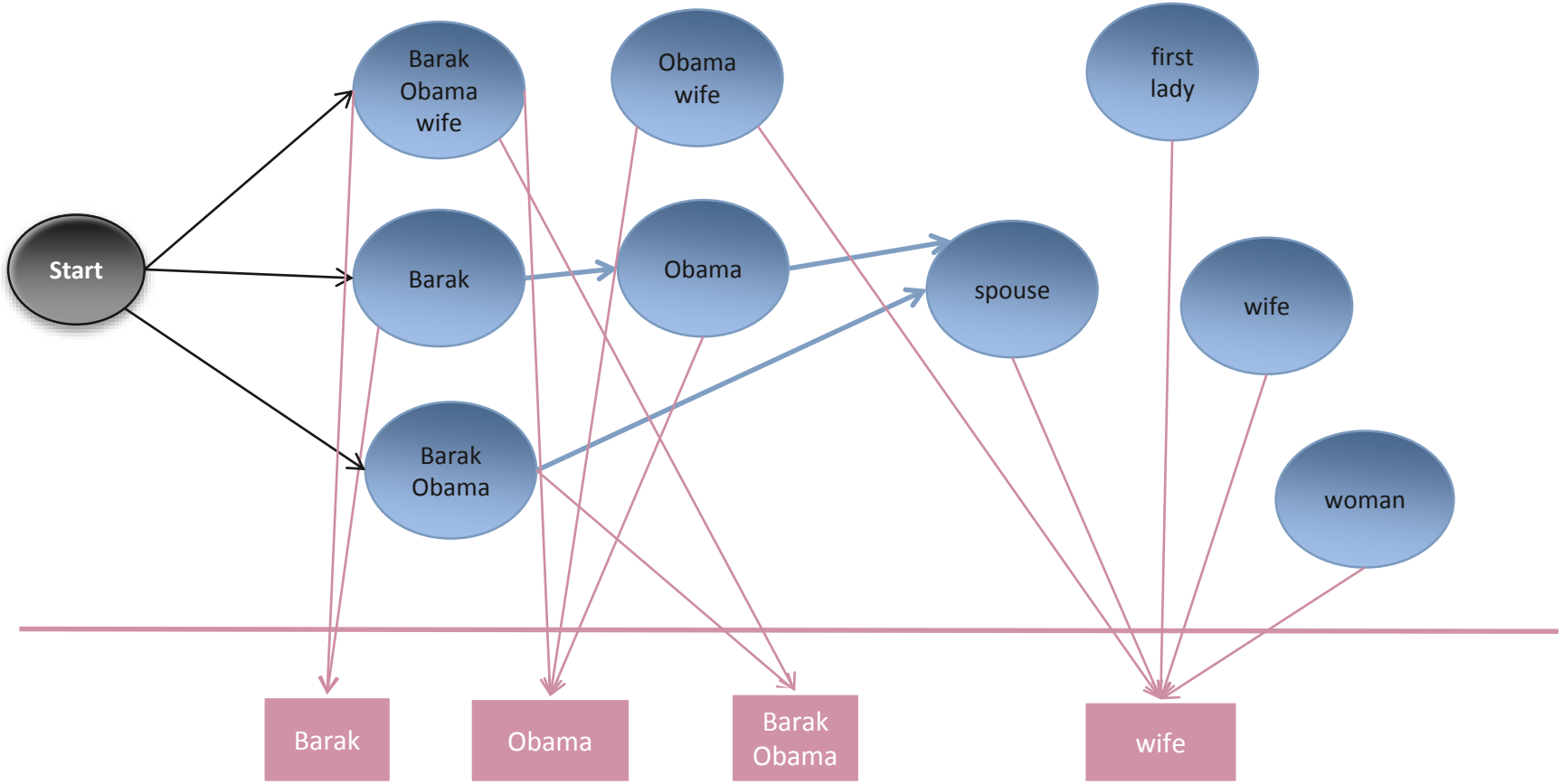


Expansion set for each segment



Reformulating query using hidden Markov model

Input query: wife of Barak Obama



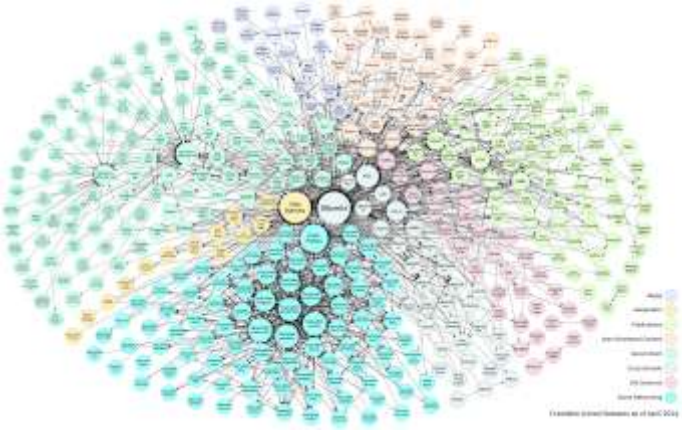
Formal Query Construction

- **Definition:** Once the resources are detected, a connected subgraph of the knowledge base graph, called the query graph, has to be determined which fully covers the set of mapped resources.

Disambiguated resources	**sider:sideEffect**
	diseasome:possibleDrug
	diseasome:1154
SPARQL query	**SELECT ?v3 WHERE {** **diseasome:115 diseasome:possibleDrug ?v1 .** **?v1 owl:sameAs ?v2 .** **?v2 sider:sideEffect ?v3 .}**

Data Fusion on Linked Data

- Answer of a question may be spread among different datasets employing heterogeneous schemas.
- Constructing a federated query from needs to exploit links between the different datasets on the schema and instance levels.



Two different approaches

Formal query construction based on

- Template-based query construction
- Forward chaining based query construction

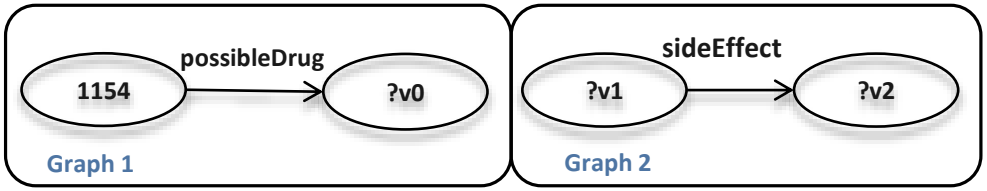
Federated query construction using forward chaining

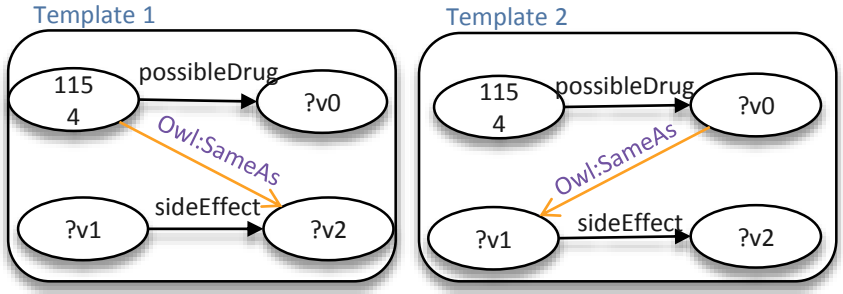
1. Set of resources

2. Incomplete query graph

3. Query graph

Query	What is the side effects of drugs used for Tuberculosis?	
resources	diseasome:1154	(type instance)
	diseasome:possibleDrug	(type property)
	sider:sideEffect	(type property)





Any question?