DEANNA: Natural Language Questions for the Web of Data

Mohamed Yahya[†] Klaus Berberich[†] Shady Elbassiousni[*]

Maya Ramanath[‡] Volker Tresp[#] Gerhard Weikum[†]

[†] Max Planck Institute for Informatics

[*] Qatar Computing Research Institute

[‡] IIT Delhi [#] Siemens

“Classical” QA

“Who played in Casablanca and was married to a writer born in Rome?”

Ingrid Bergman (29 August 1915 – 29 Augu
a Swedish actress who starred in a variety of
and American films. She won three Academy
Awards, and the Tony Award for Best Actress
the fourth greatest female star of American
by the American Film Institute. She is best re
roles as Ilsa Lund in *Casablanca* (1942),
II dram
in *Noto*
Grant.

Rossellini was born in Rome. His mother, Elettra (née Bellan), was a housewife, and his father, Angiolo Giuseppe "Beppino" Rossellini, owned a construction firm. His mother was of part French descent, from immigrants who had arrived in Italy during the Napoleonic Wars. He lived on the Via Ludovisi, whereBenito Mussolini had his first Roman hotel in 1922 when Fascism obtained power in Italy.

In 1950, after a decade of stardom i
the Italian film *Stromboli*, which led
director Roberto Rossellini while they were both already married. The affair and then **marriage with Rossellini** created a scandal that forced her to remain in Europe until 1956, when she made a successful Hollywood r
Academy Av
personal an
University C

Roberto Gastone Zeffiro Rossellini (8 May 1906 – 3 June 1977) was an Italian film director and **screenwriter**. Rossellini was one of the directors of the Italian neorealist cinema, contributing films such as *Roma città aperta* (*Rome, Open City* 1945) to the movement.

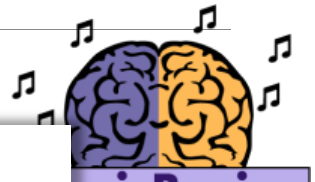
QA, meet the (semantic) Web of Data

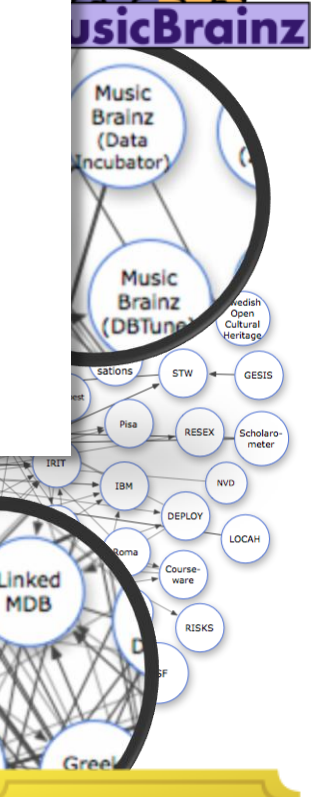
WordNet

+

LOD: > 31 BILLION triples

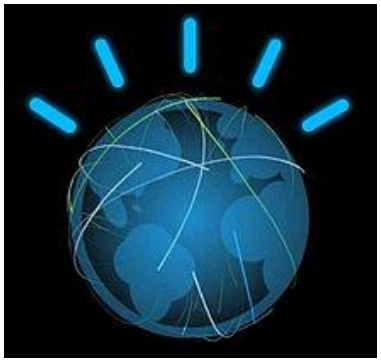
YAGO2: 120 million facts on 10 million entities





Subject	Predicate	Object
Rome	isA	city
city	subclassOf	location
Roberto_Rossellini	mariedTo	Ingrid_Bergman

Semantics Types



U.S. Cities

ITS LARGEST AIRPORT WAS NAMED FOR A WORLD WAR II HERO; ITS SECOND LARGEST, FOR A WORLD WAR II BATTLE

What is Toronto???

Crash Course: Querying the Web of Data

"Who played in Casablanca and was married to a writer born in Rome?"

Projection Variable

SPARQL

```
SELECT ?p WHERE{
   ?p type person .
   ?p actedIn Casablanca_(film).
   ?p isMarriedTo ?w .
   ?w type writer .
   ?w bornIn Rome }
```

- Dot: conjunction
- ?p, ?w: variables to be bound
- Same variable → same binding (join)

What is DEANNA?

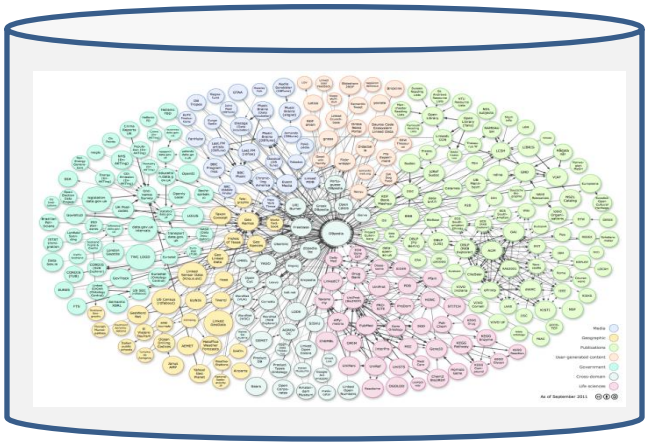
Question

"Who played in Casablanca and was married to a writer born in Rome?"

DEANNA

SPARQL

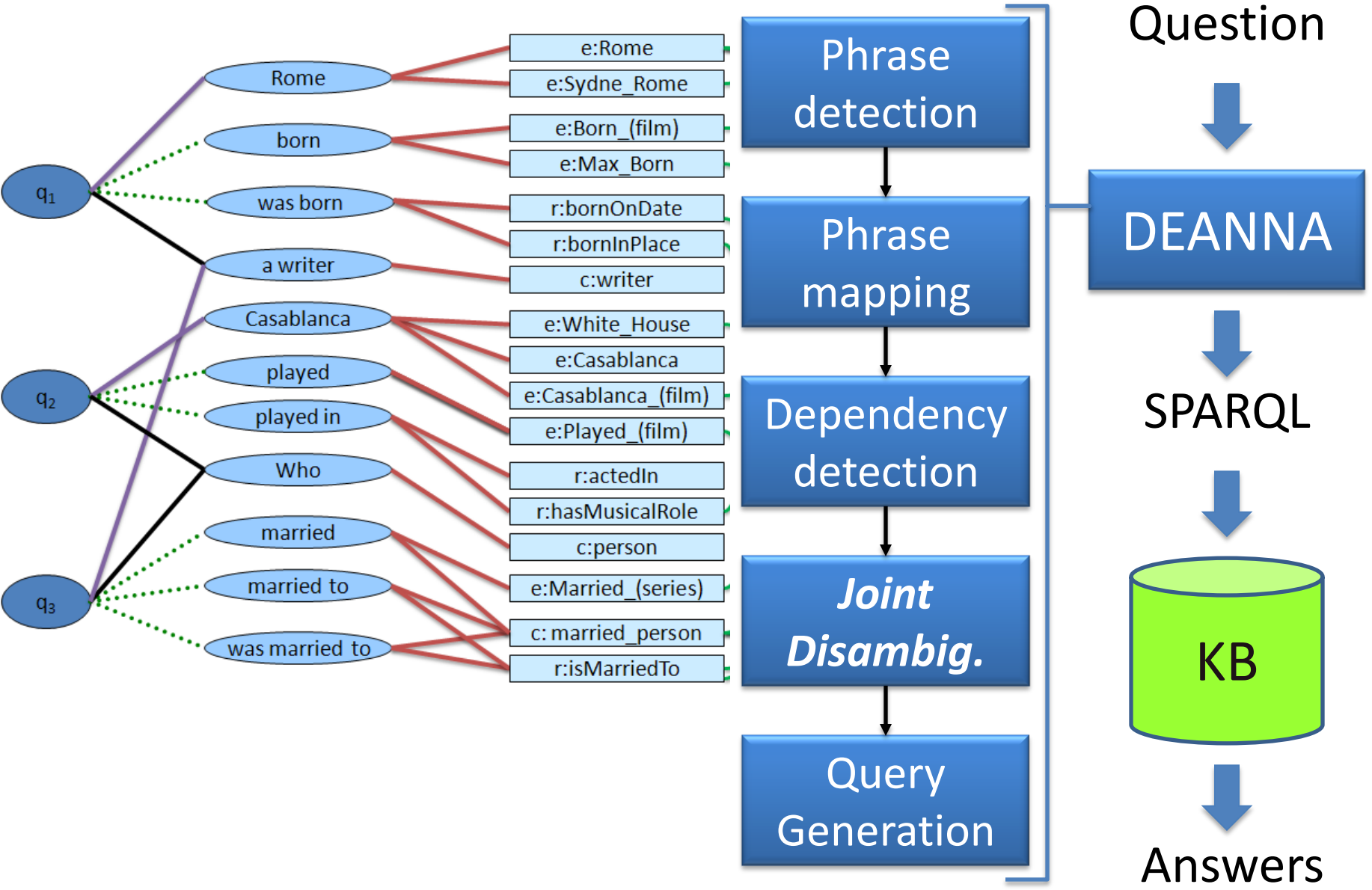
```
?p type person.
?p actedIn Casablanca_(film).
?p isMarriedTo ?w.
?w type writer .
?w bornIn Rome .
```



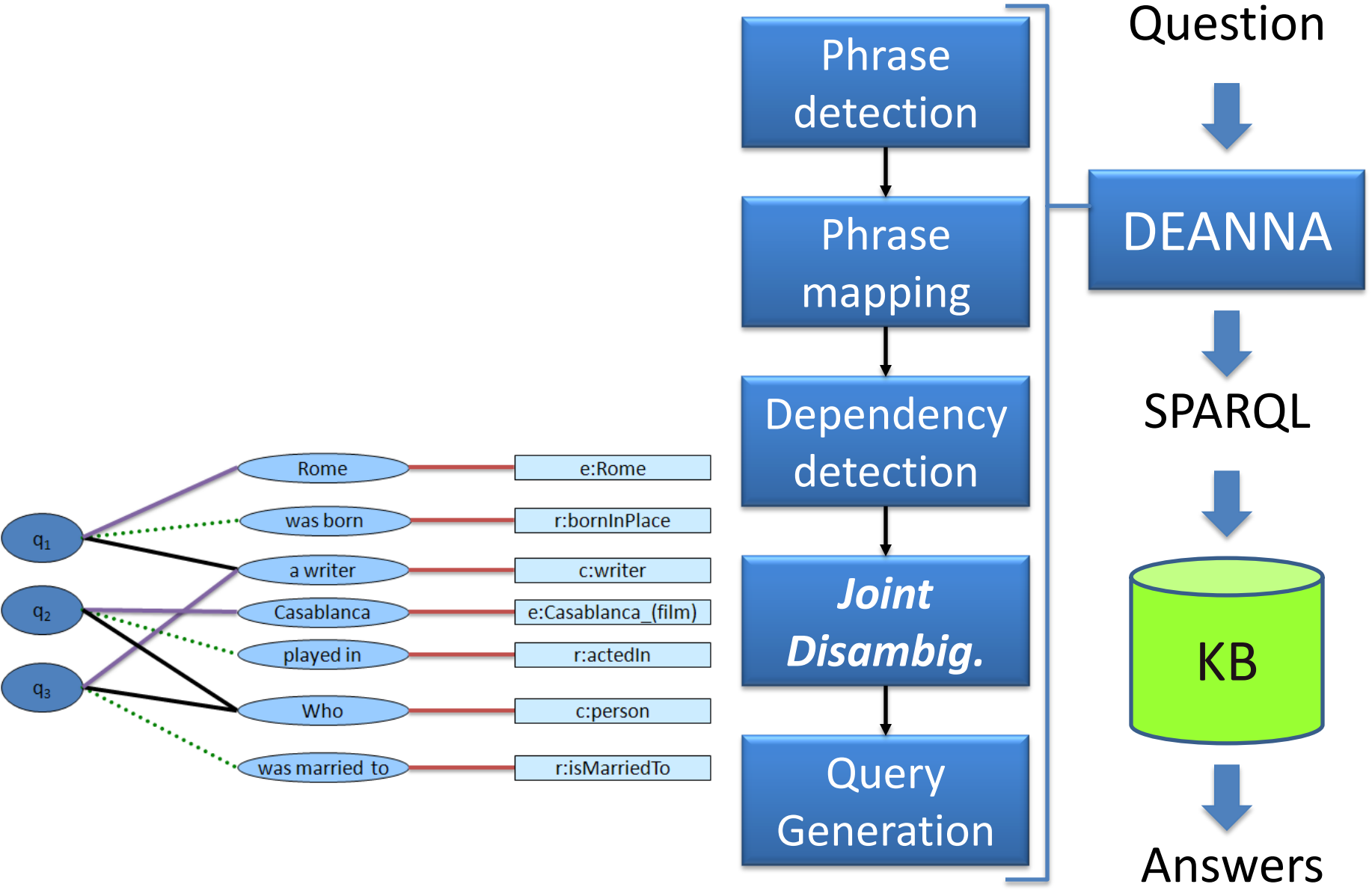
KB

Answers

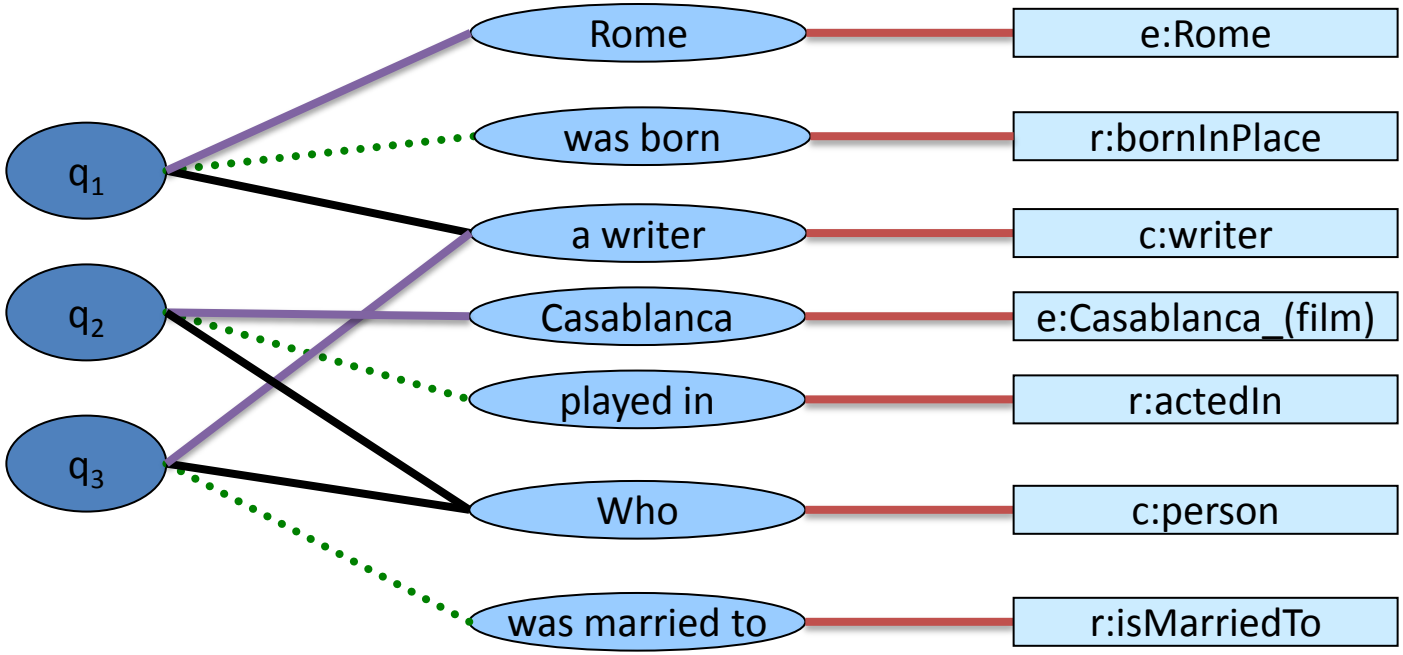
Inside DEANNA



Inside DEANNA

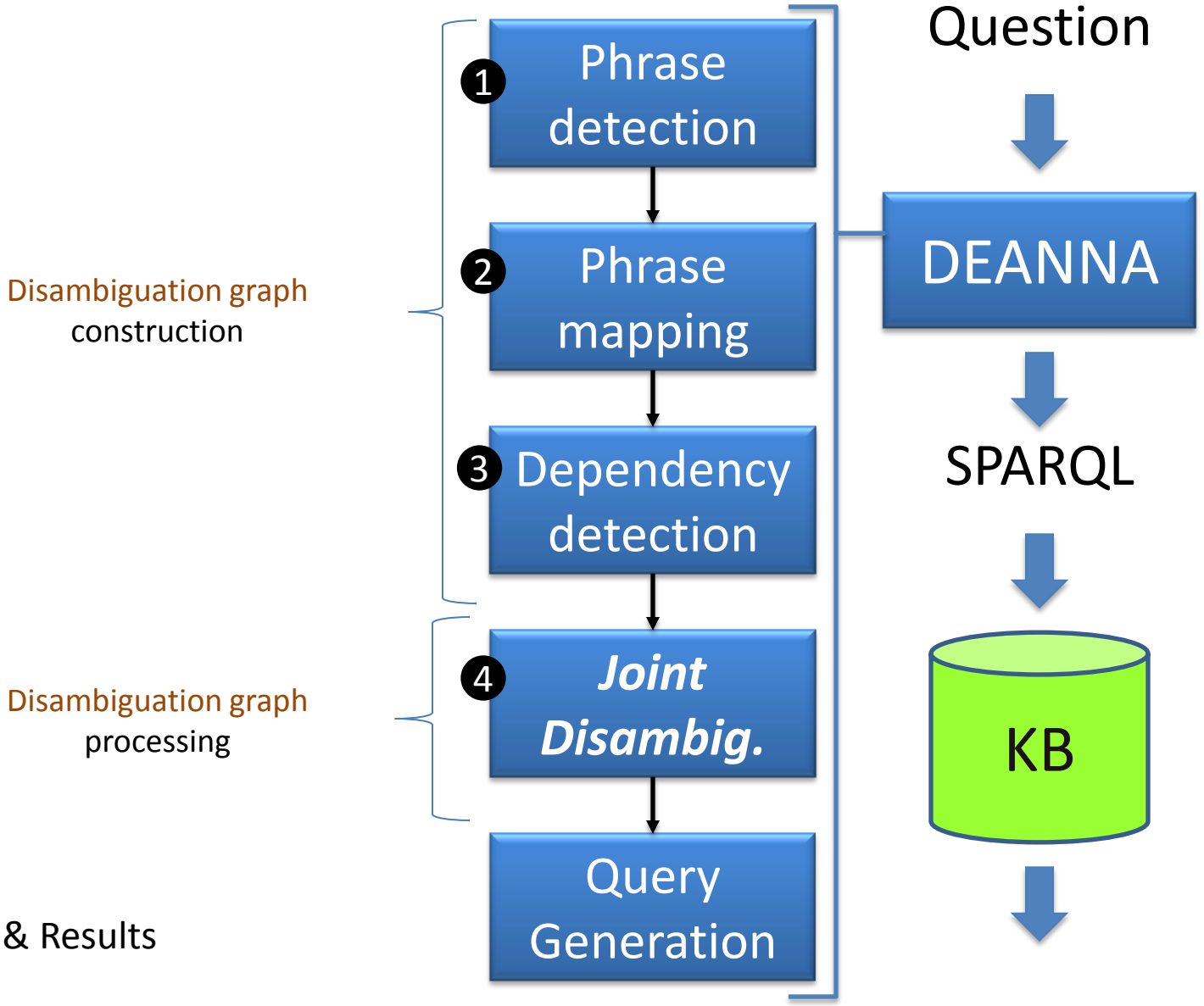


Structured Query Generation



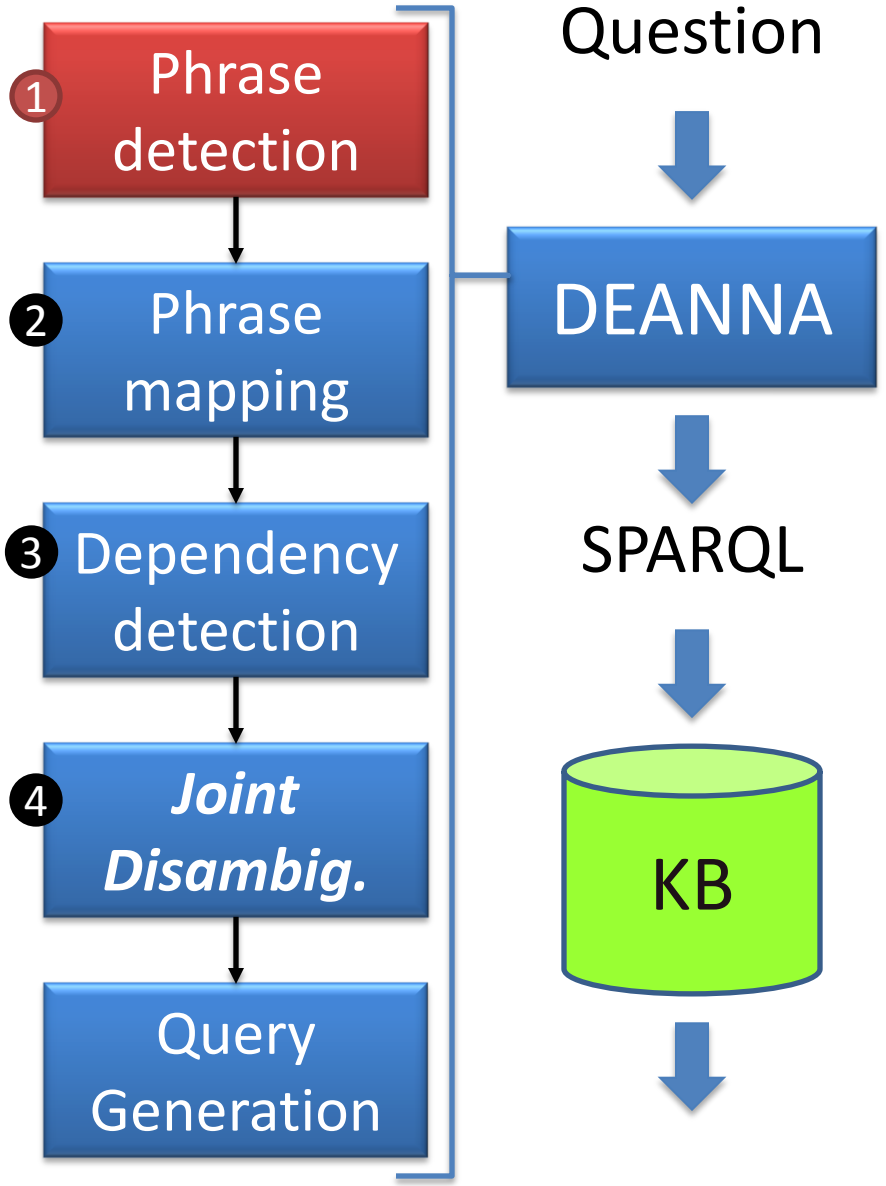

```
SELECT ?p WHERE {
     ?p type person.
     ?p actedIn Casablanca_(film).
     ?p isMarriedTo ?w.
     ?w type writer .
     ?w bornIn Rome }
```

Outline



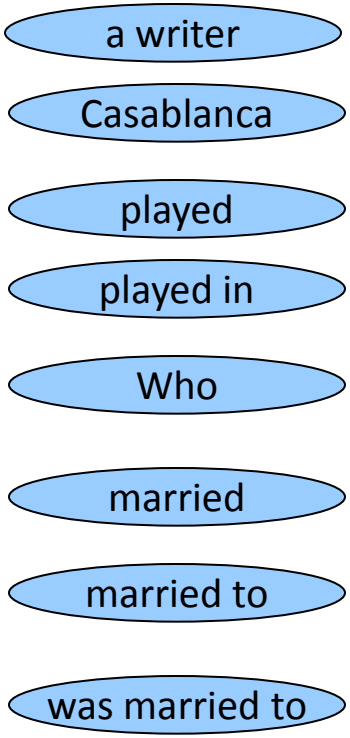
5 Experiments & Results

Outline



5 Experiments & Results

Phrase Detection



- **Concepts: entities & classes:**

Dictionary-based

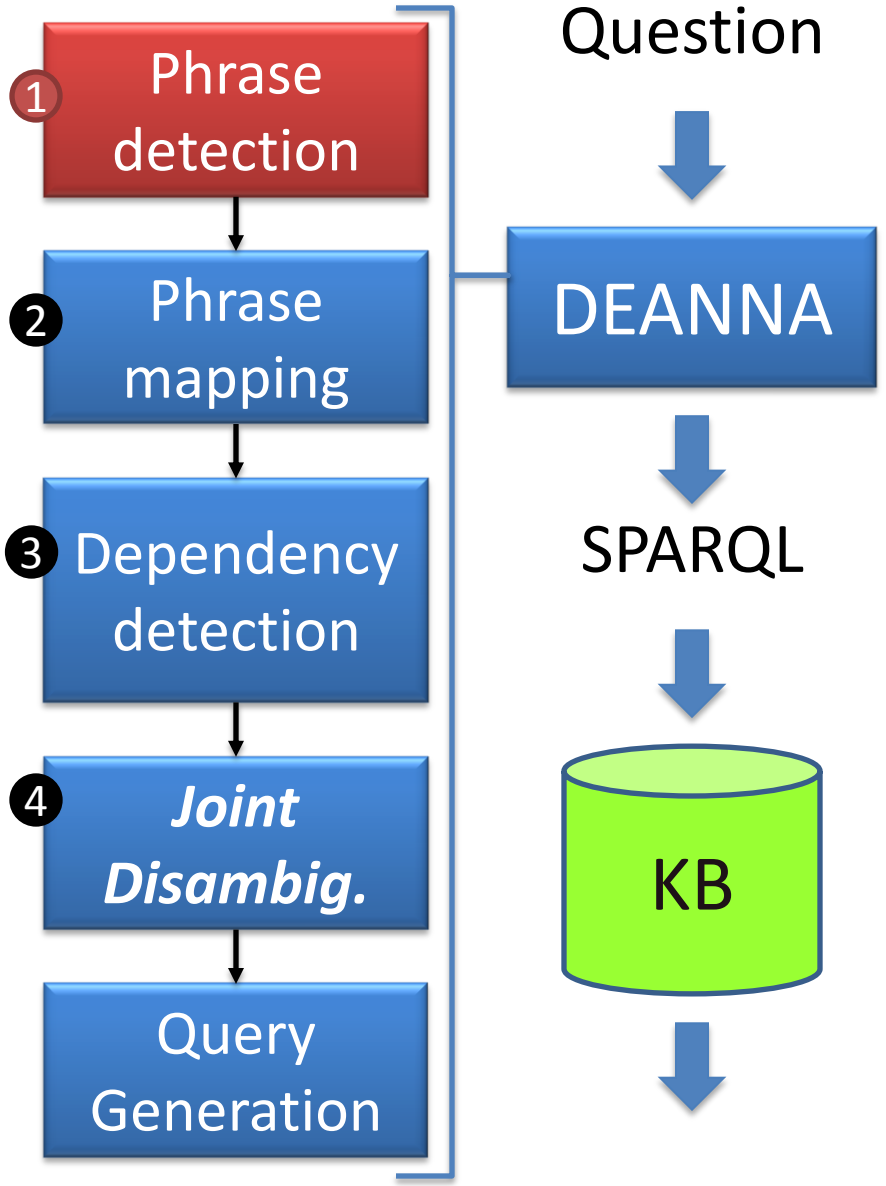
Concept	Phrase
`Casablanca`	Casablanca
`Casablanca`	Casablanca, Morocco
`Casablanca_(film)`	Casablanca the film
`Casablanca_(film)`	Casablanca
…	…

- **Relations:**

Mainly use Reverb [Fader et al. EMNLP'11]: `V | VP | VW*P`

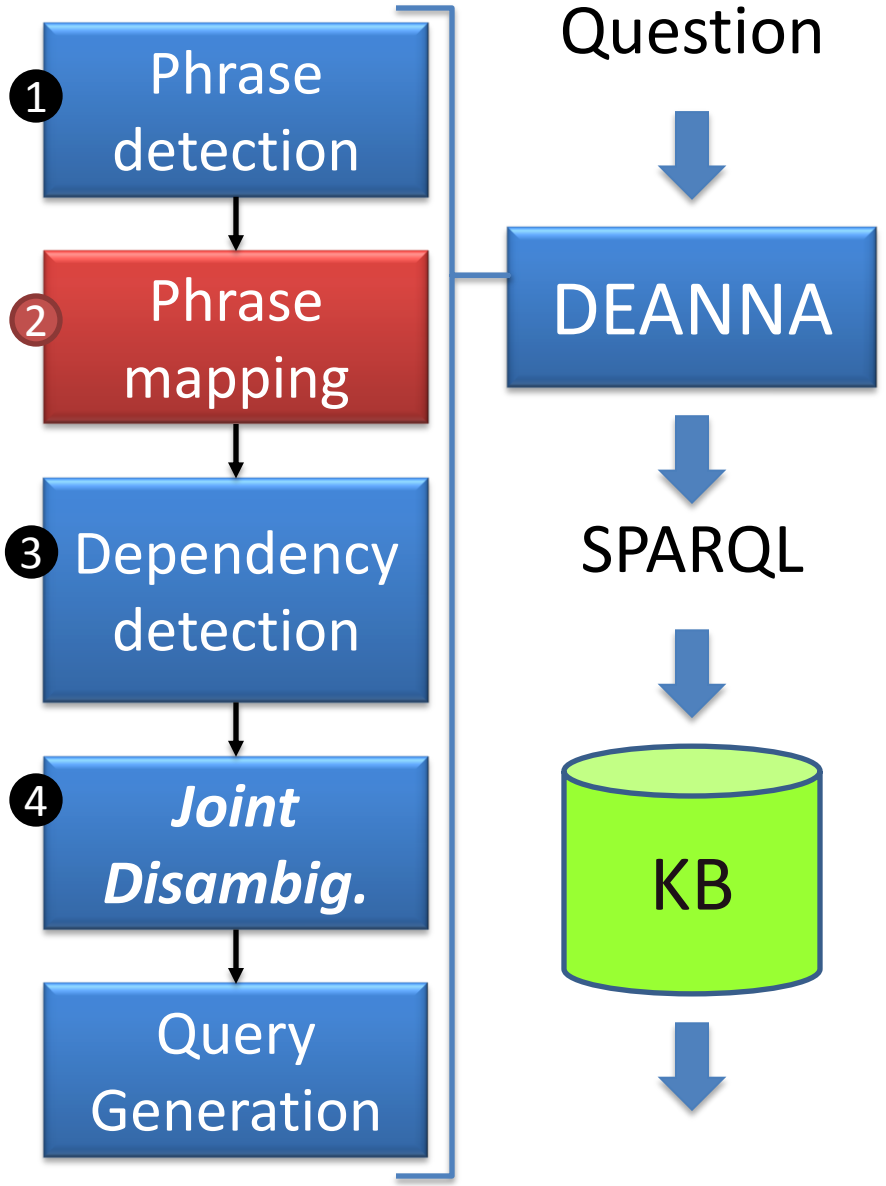
… was/VBD married/VBN to/TO a/DT…

Outline



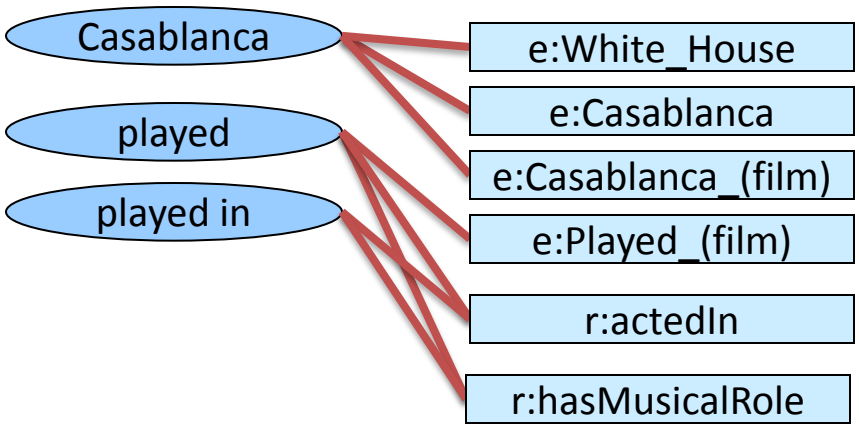
5 Experiments & Results

Outline



5 Experiments & Results

Phrase Mapping



- **Concepts: entities & classes:**

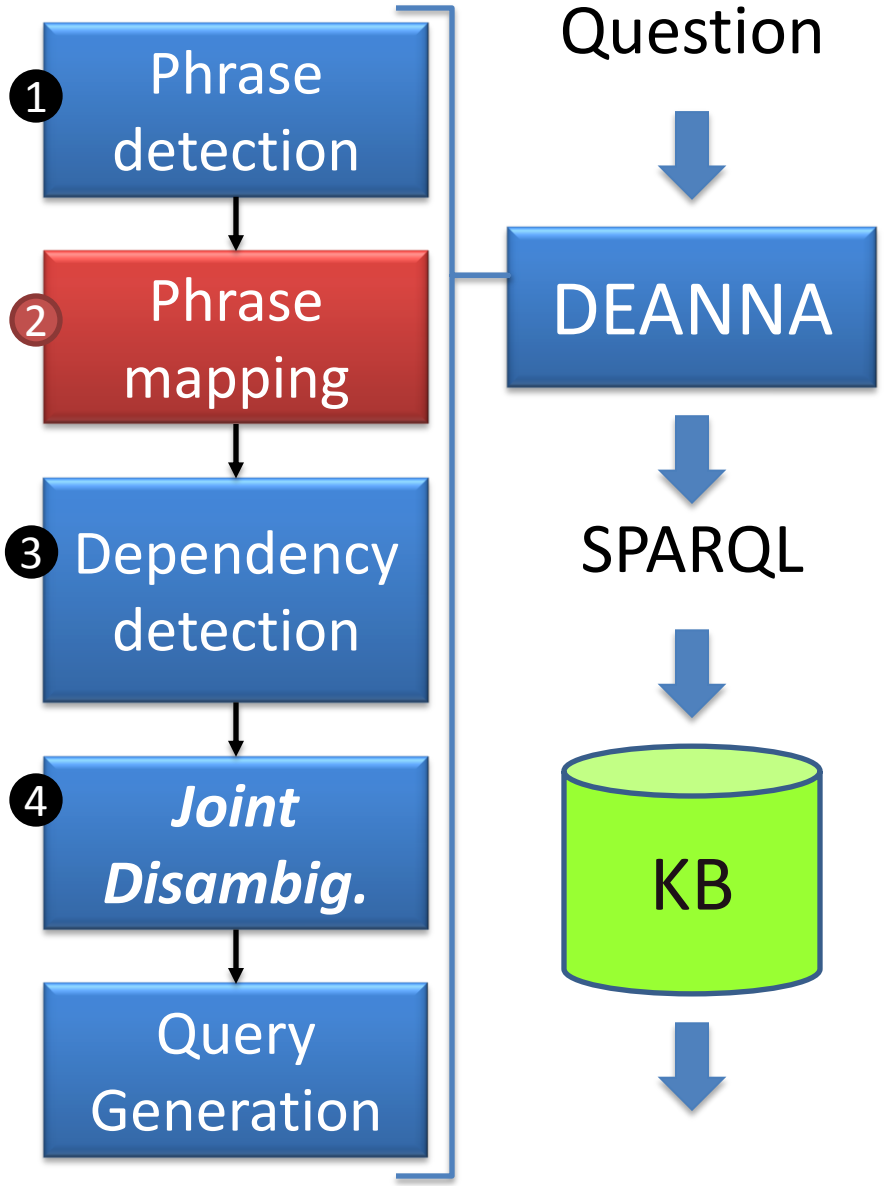
Dictionary-based

Concept	Phrase
`Casablanca`	Casablanca
`Casablanca`	Casablanca, Morocco
`Casablanca_(film)`	Casablanca the film
`Casablanca_(film)`	Casablanca
`Played_(film)`	Played

- **Relations: Dictionary -based**

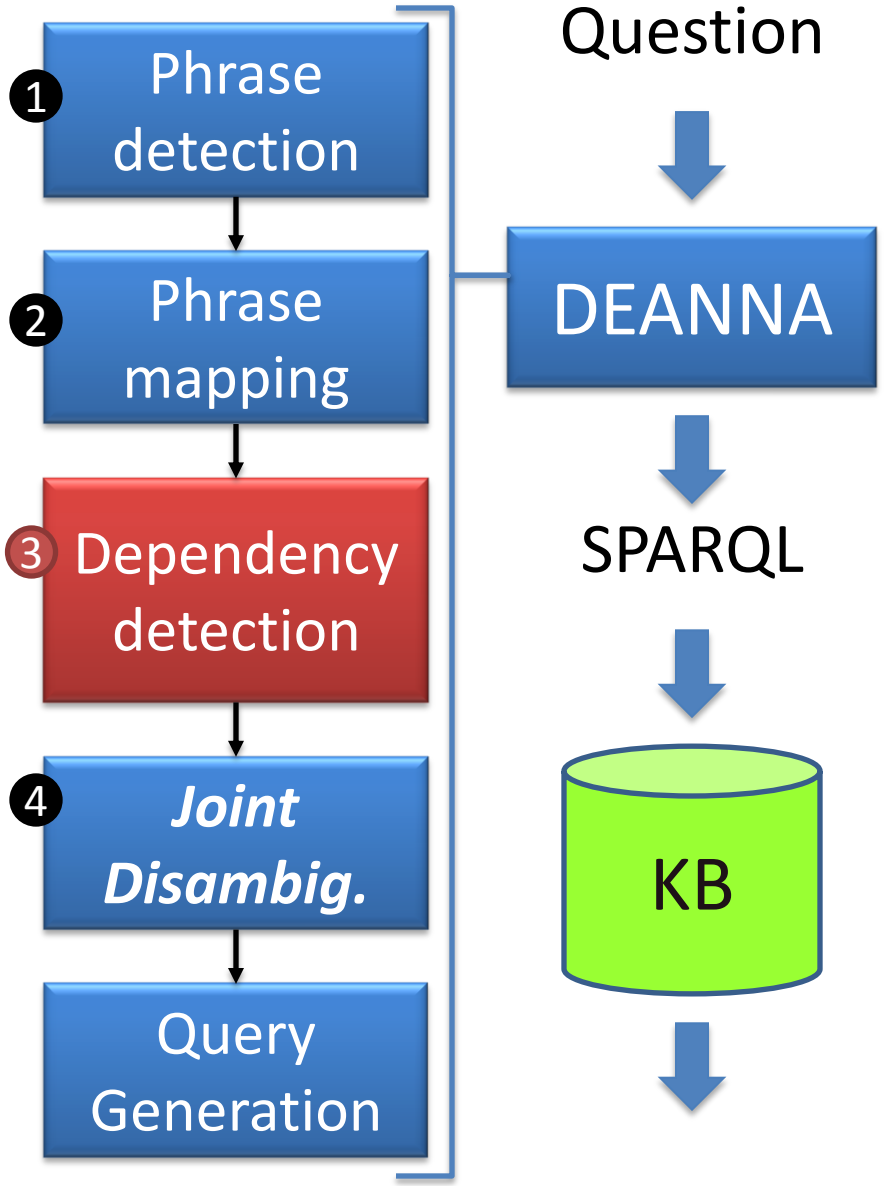
Relation	Phrase
`actedIn`	acted in
`actedIn`	played in
`hasMusicalRole`	plays
`hasMusicalRole`	mastered

Outline



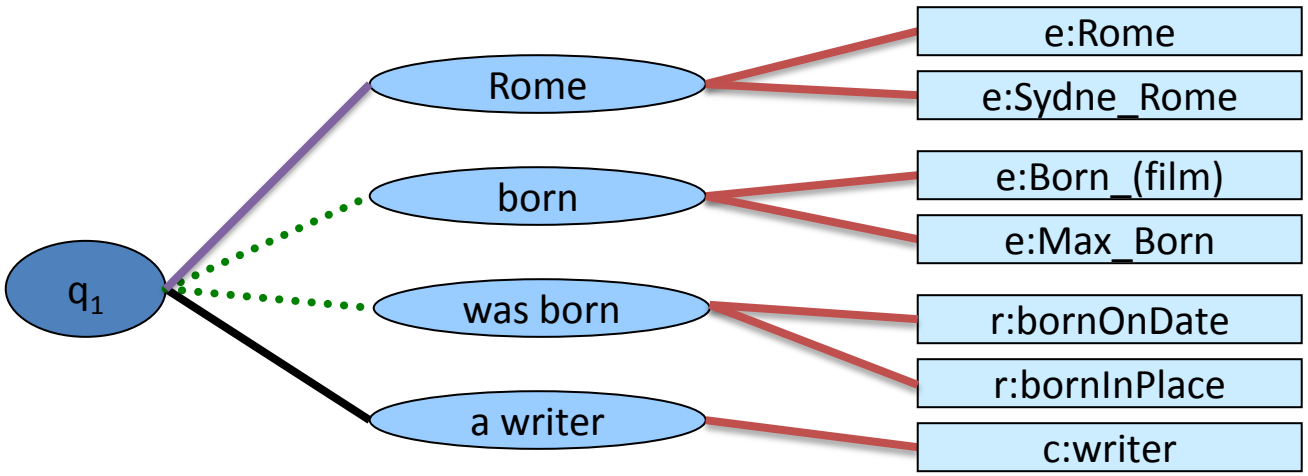
5 Experiments & Results

Outline

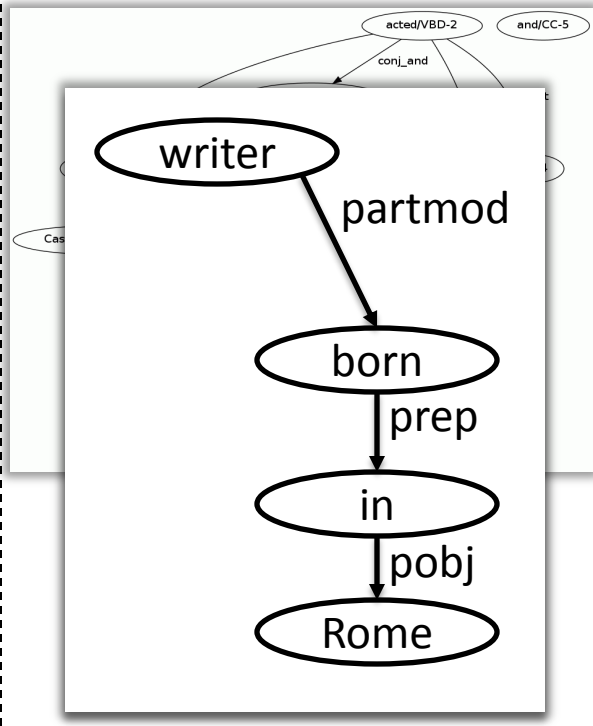


5 Experiments & Results

Dependency Detection

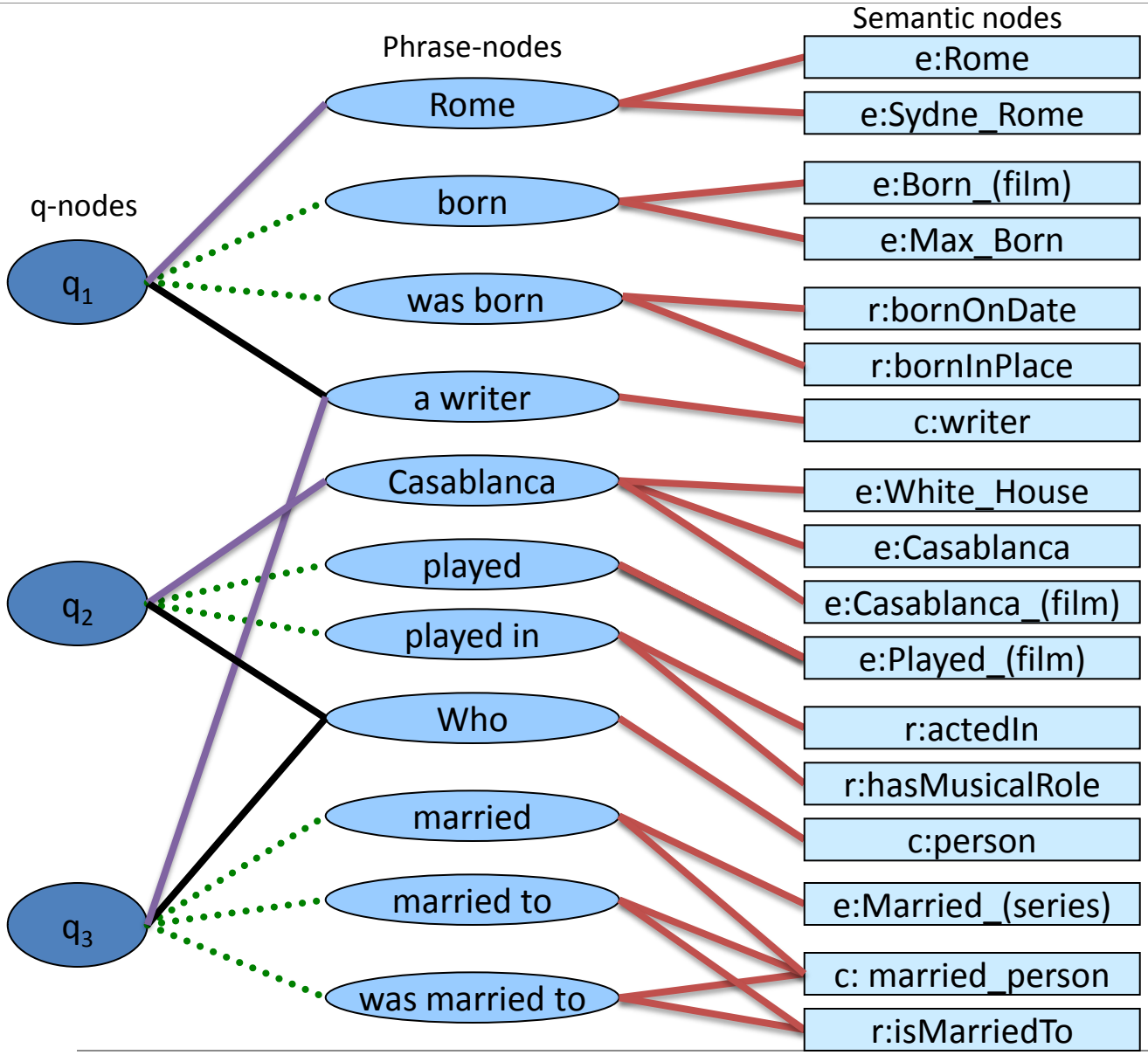


Look for specific patterns in dependency parses [de Marneffe et al. LREC'06]

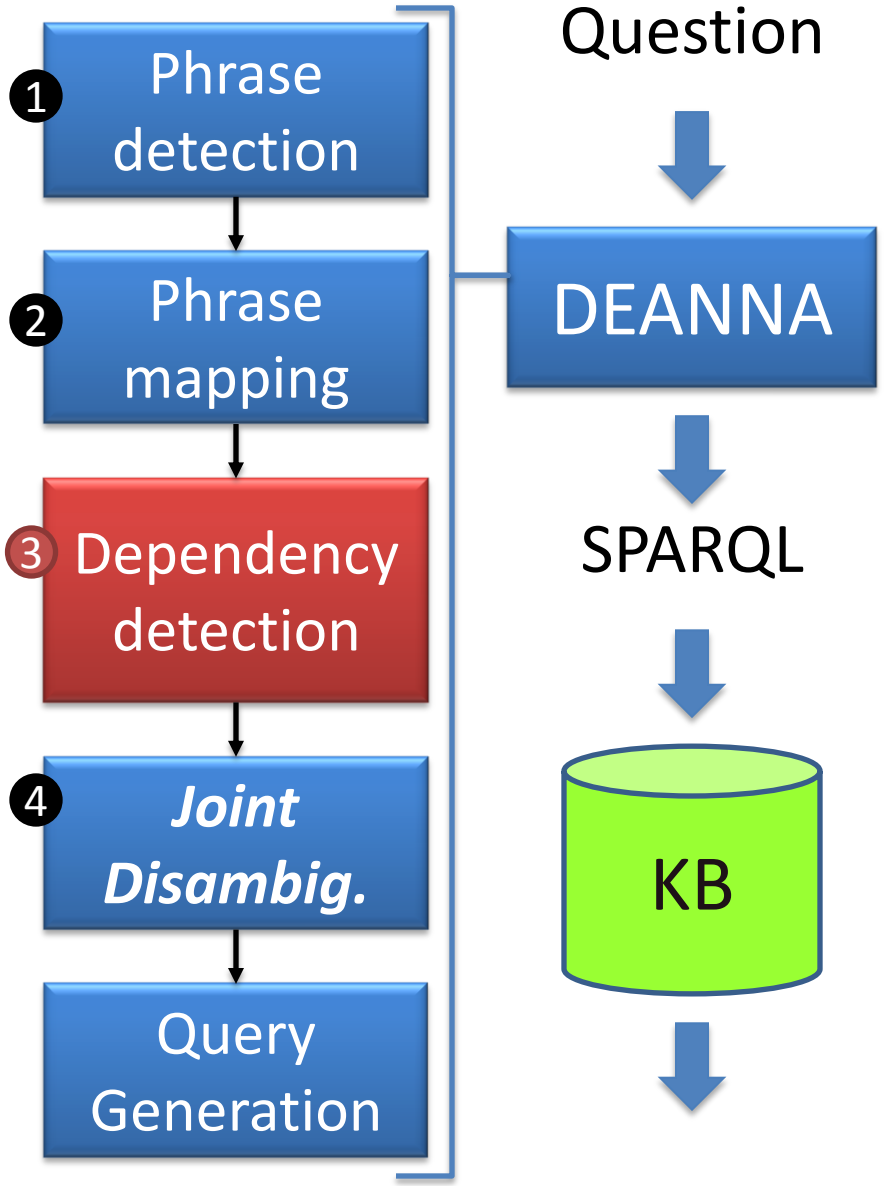


RESULT

Disambiguation Graph

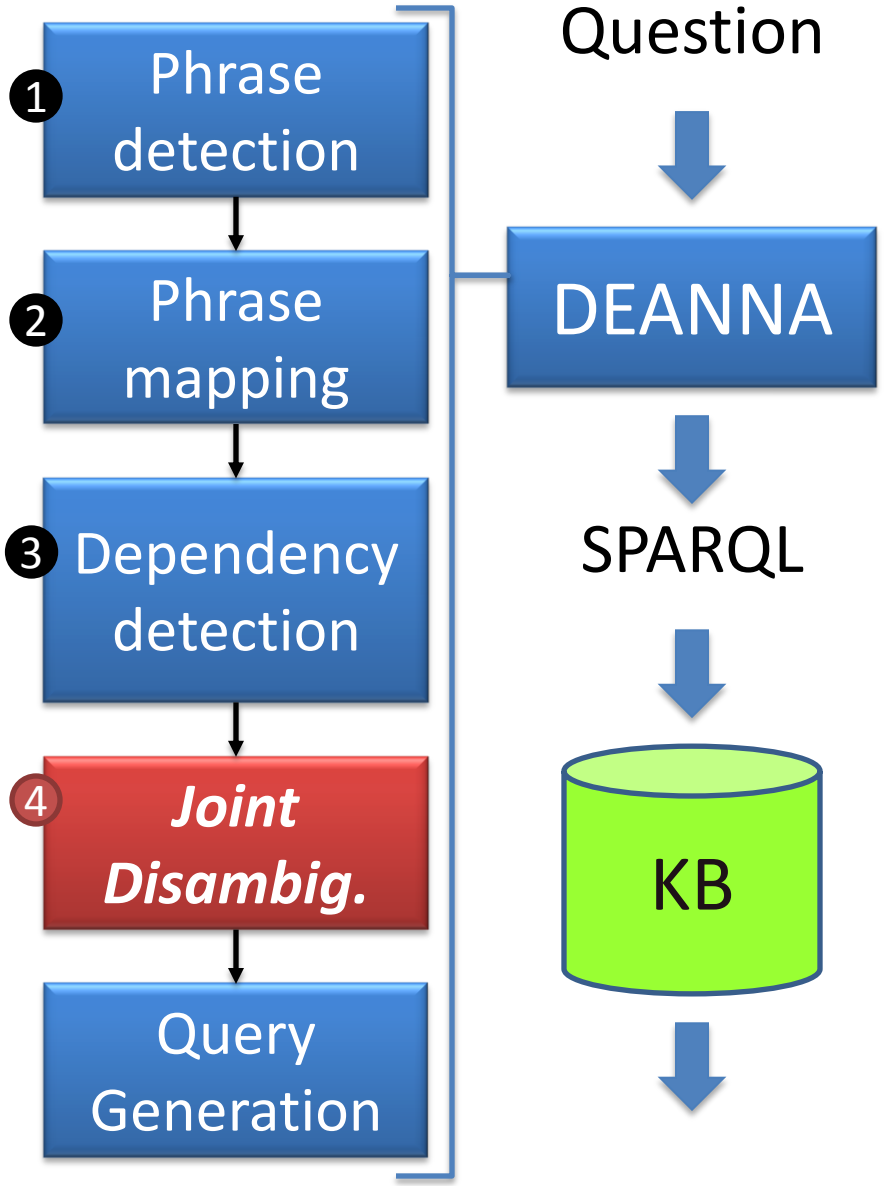


Outline



5 Experiments & Results

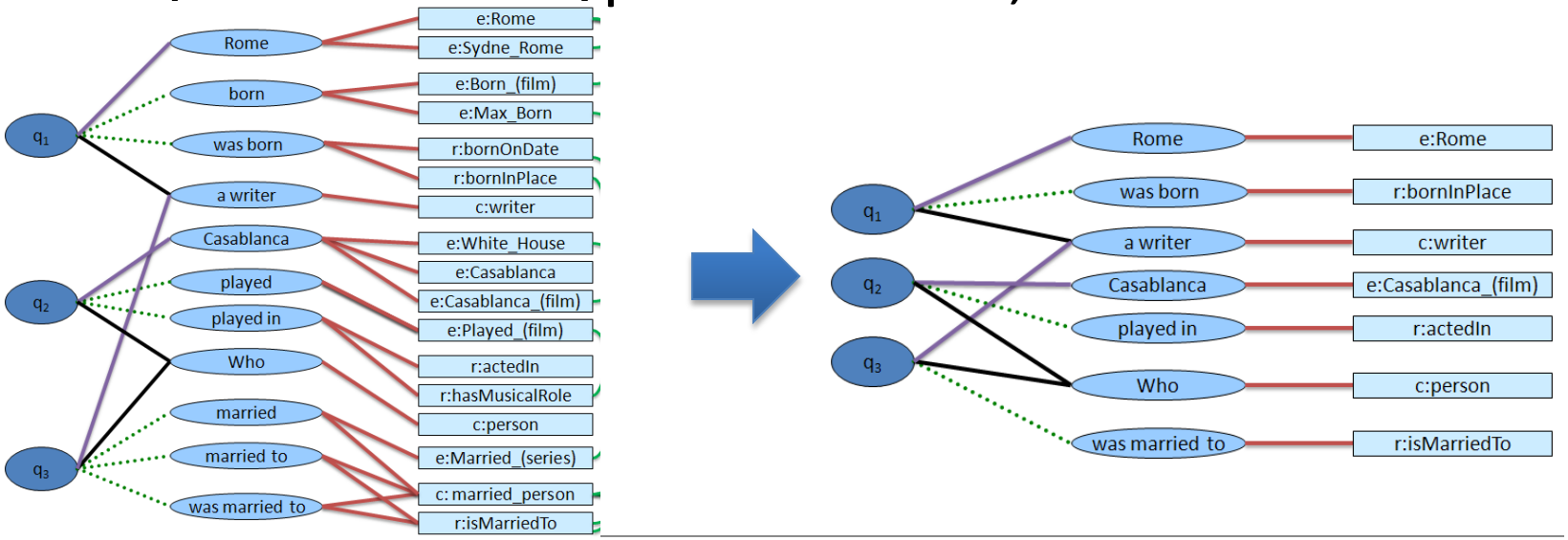
Outline



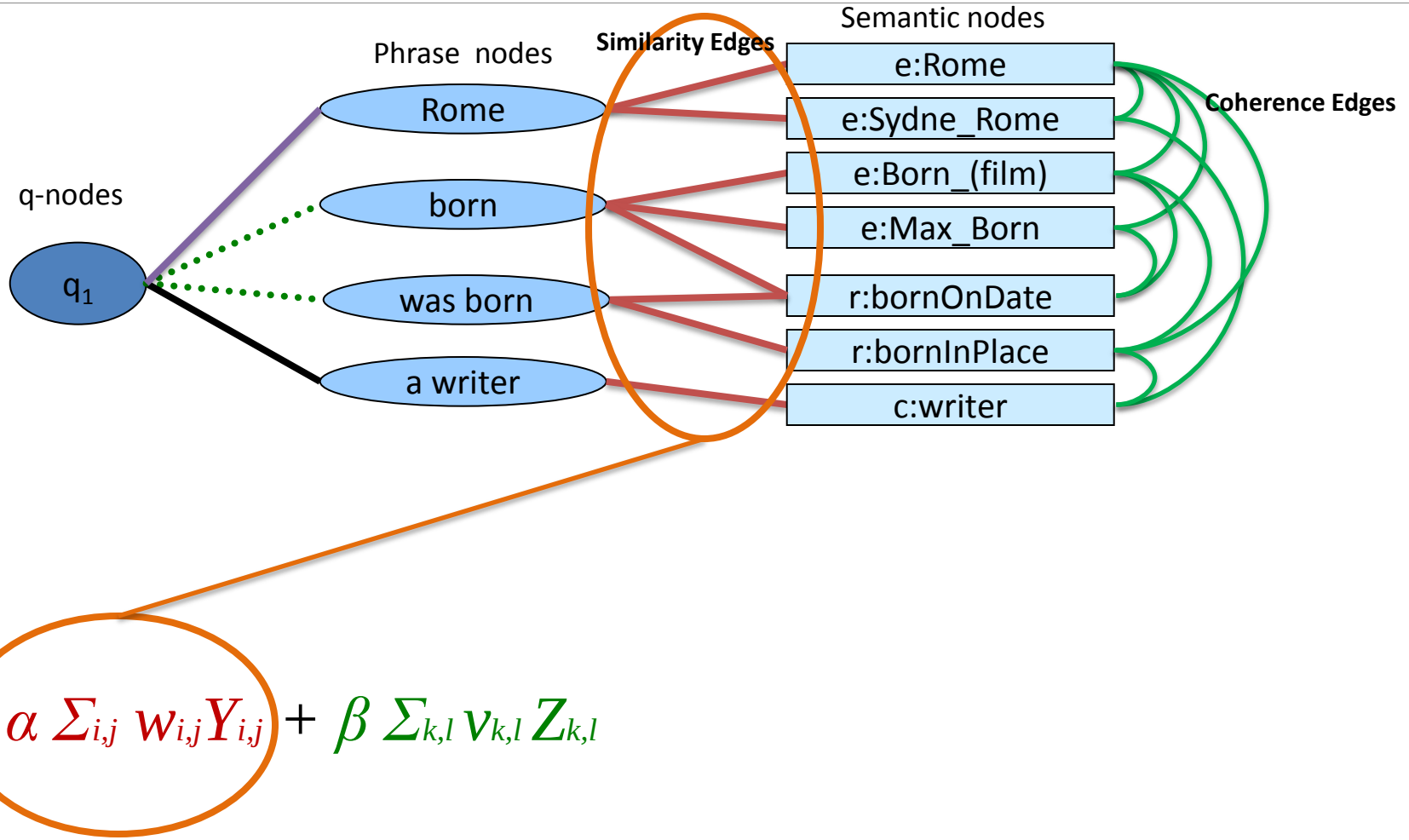
5 Experiments & Results

Joint Disambiguation - ILP

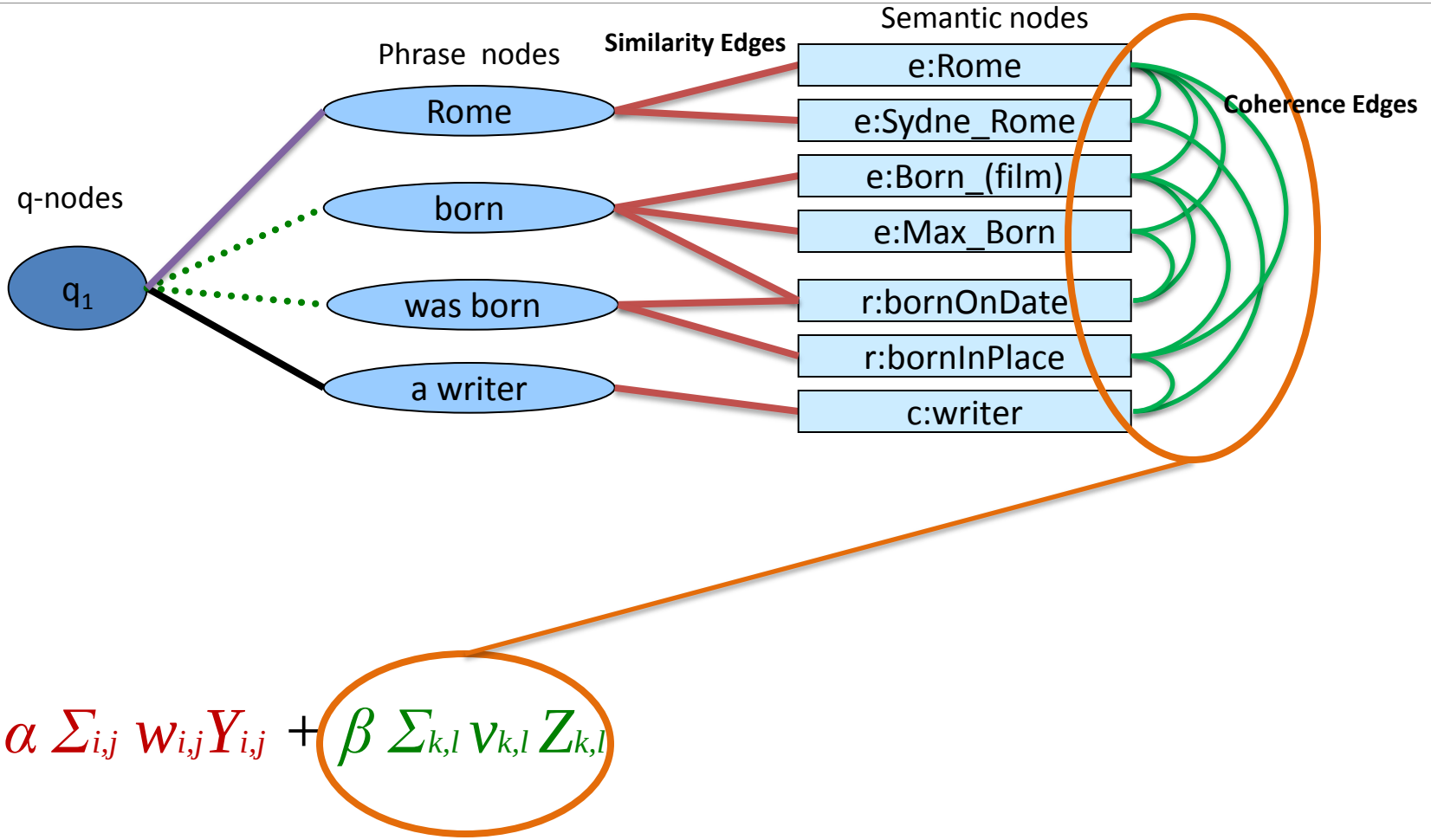
- *ILP:* Integer Linear Programming
- *maximize* $\alpha \sum_{i,j} w_{i,j} Y_{i,j} + \beta \sum_{k,l} v_{k,l} Z_{k,l} + \ldots$
- Subject to:
  - No token in multiple phrases,
  - Triples observe type constraints, …



Joint Disambiguation – Objective

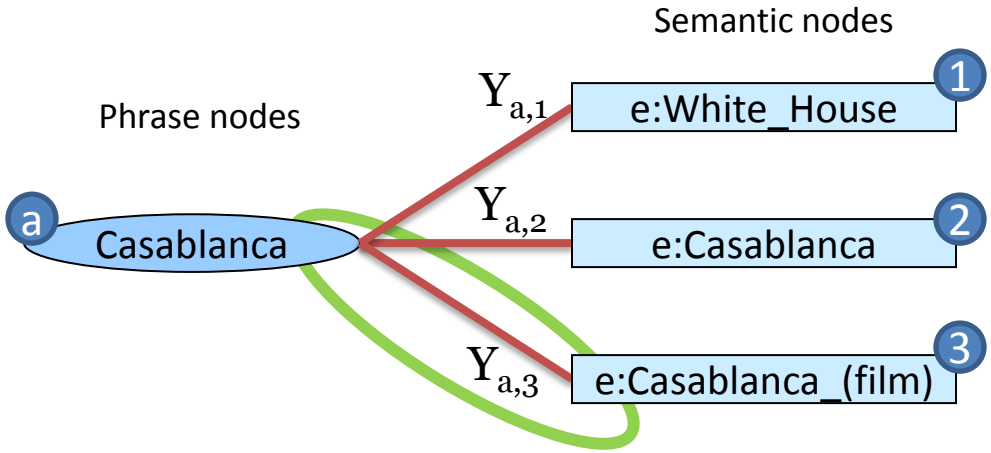


Joint Disambiguation – Objective



Joint Disambiguation – Constraints

A phrase node can be assigned to one semantic node:



$$\alpha \sum_{i,j} w_{i,j} Y_{i,j} + \beta \sum_{k,l} v_{k,l} Z_{k,l}$$

Joint Disambiguation – Constraints

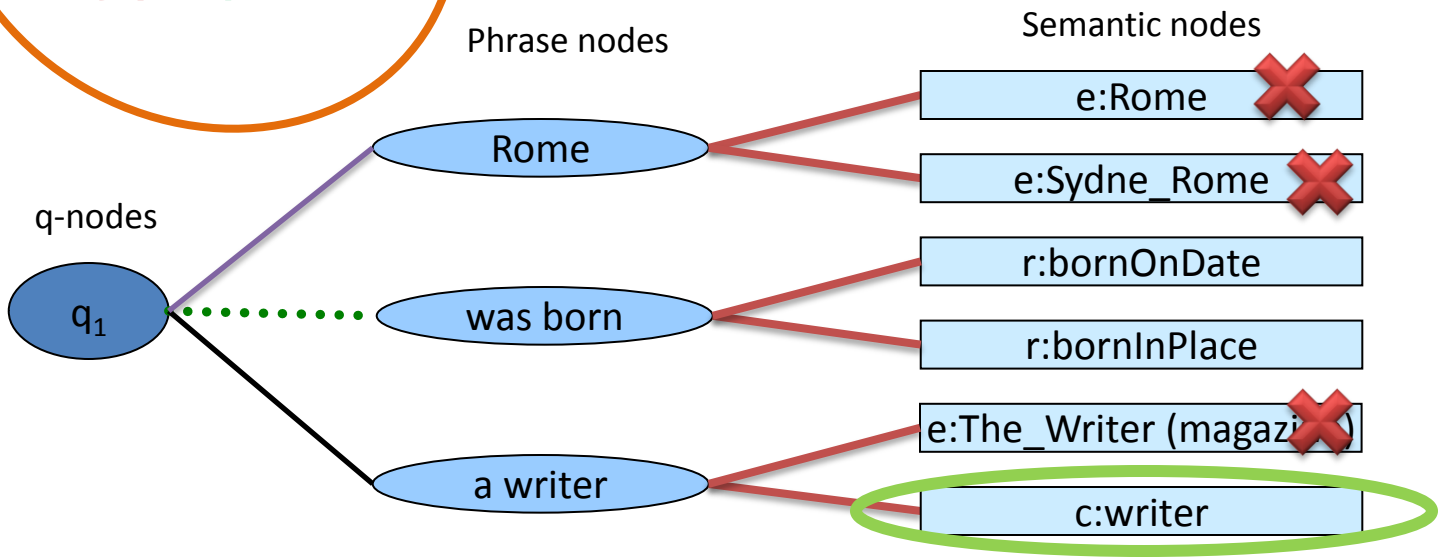
Classes translate to type-constrained variables

→ Every semantic triple should have a class to *join* & *project*!

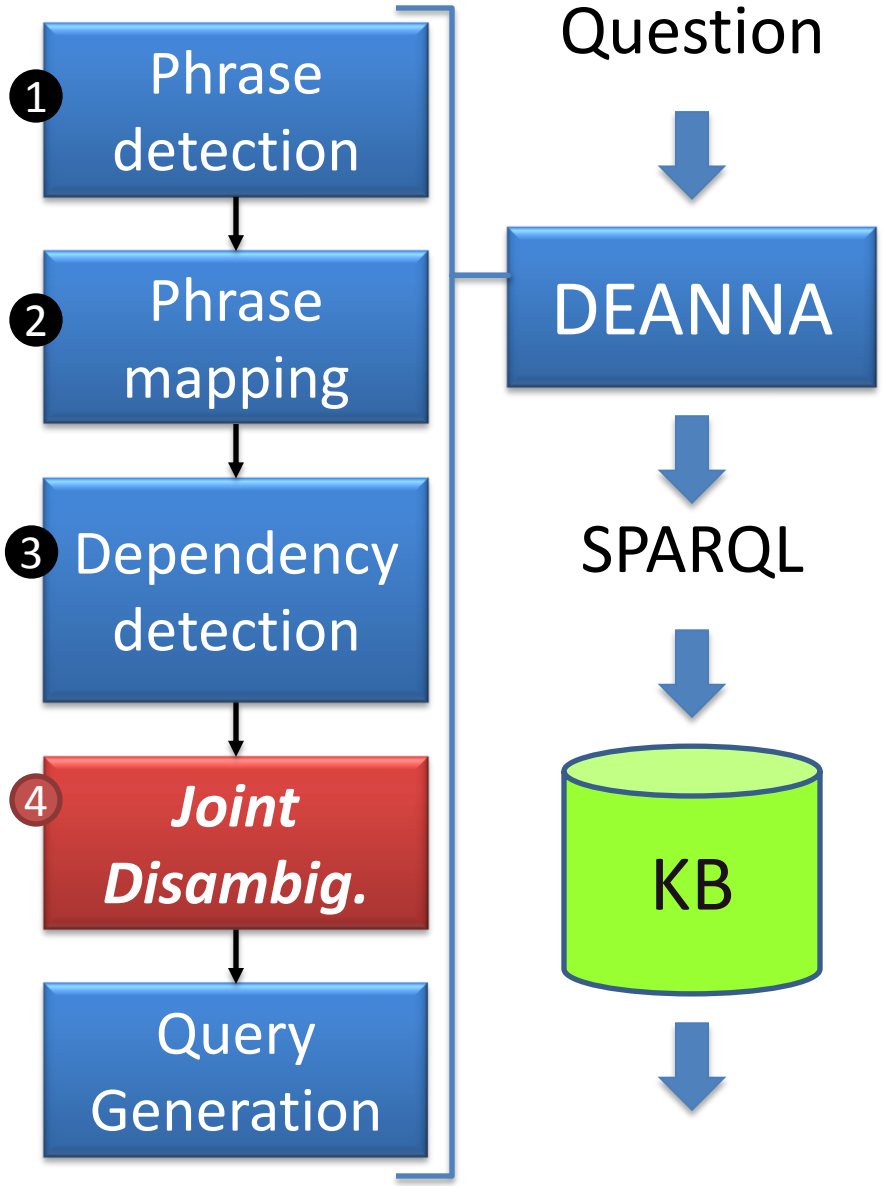
```
person actedIn Casablanca_(film)
```

▼

```
?x type person . ?x actedIn Casablanca_(film)
```

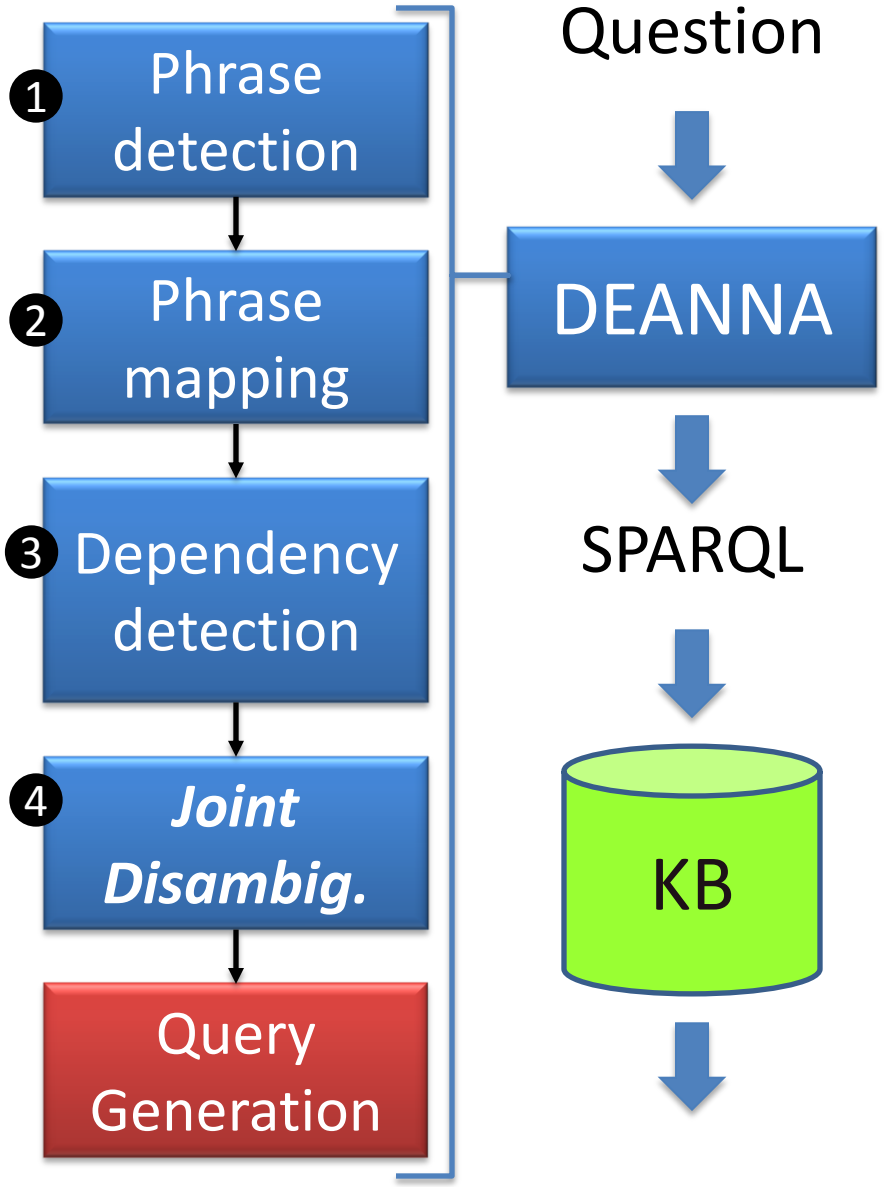


Outline



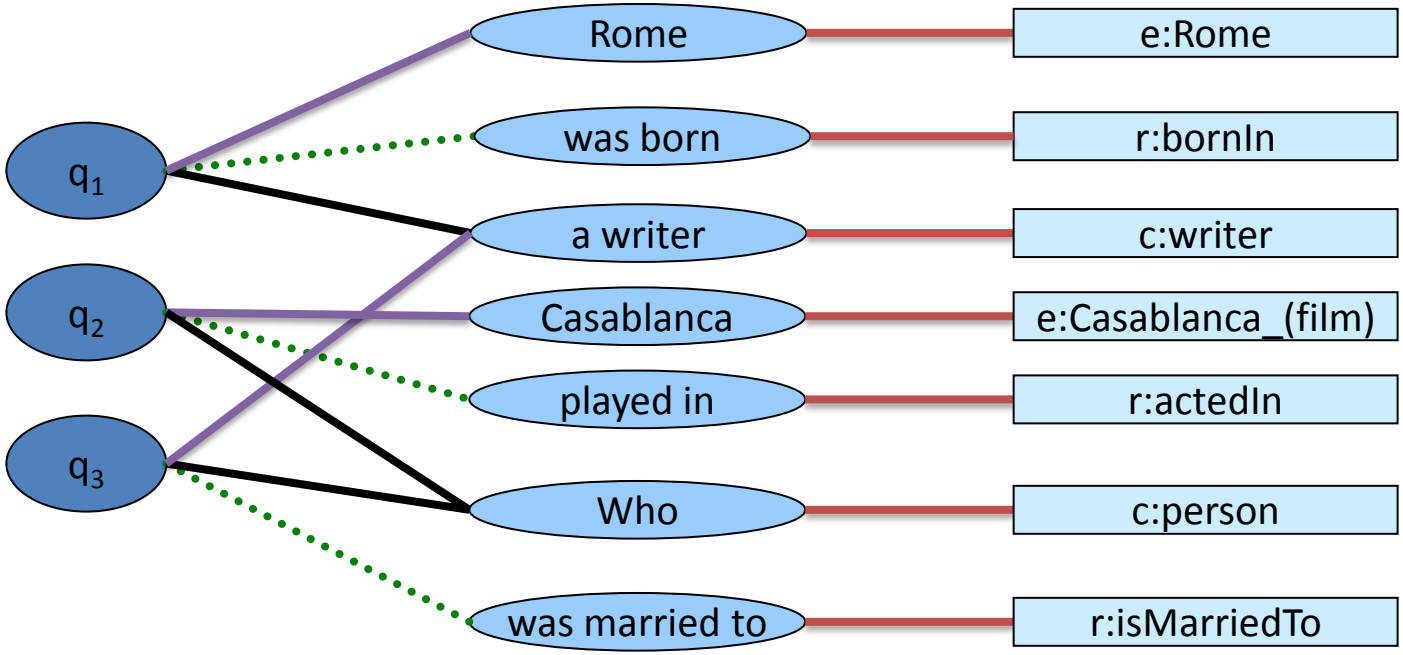
5 Experiments & Results

Outline



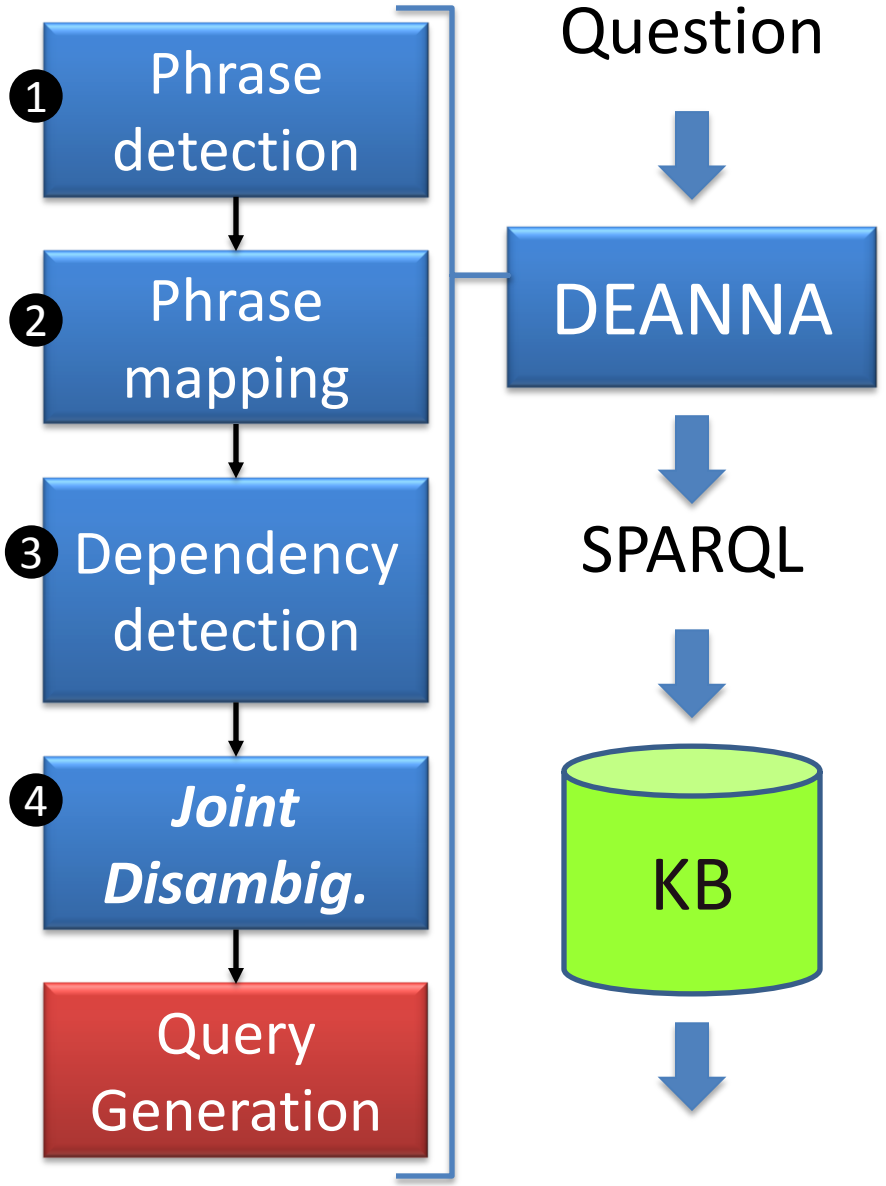
5 Experiments & Results

Structured Query Generation



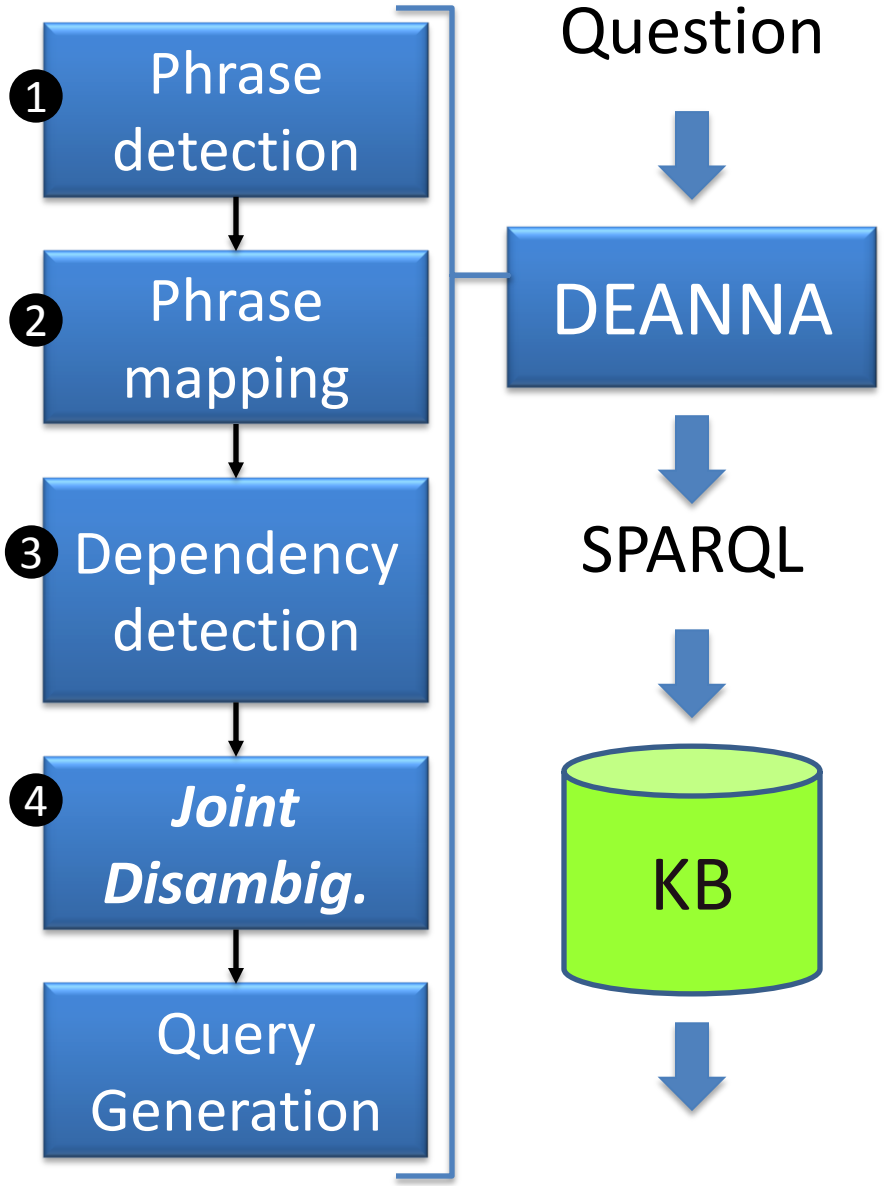

```
SELECT ?p WHERE {
     ?w type writer .
     ?w bornIn Rome .
     ?p type person.
     ?p actedIn Casablanca_(film).
     ?p isMarriedTo ?w }
```

Outline



5 Experiments & Results

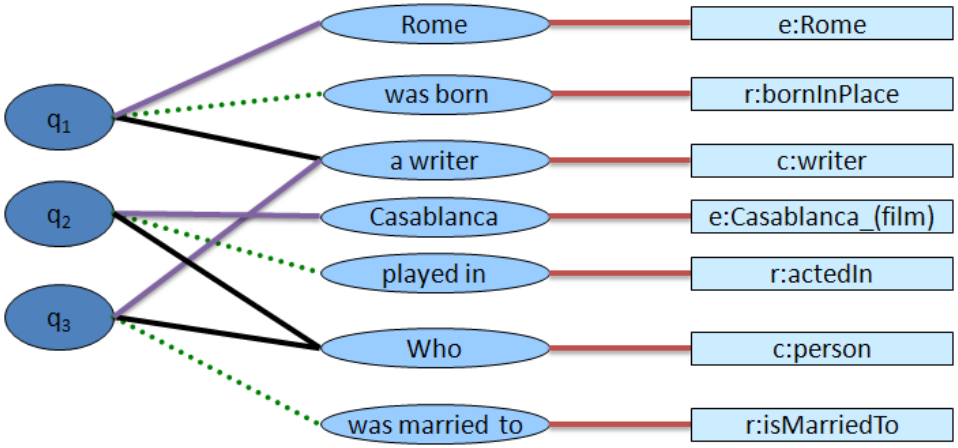
Outline



5 Experiments & Results

Evaluation Methodology

- 3-stage evaluation for more insight:
  1. Disambiguation
  2. Query generation
  3. Query answering

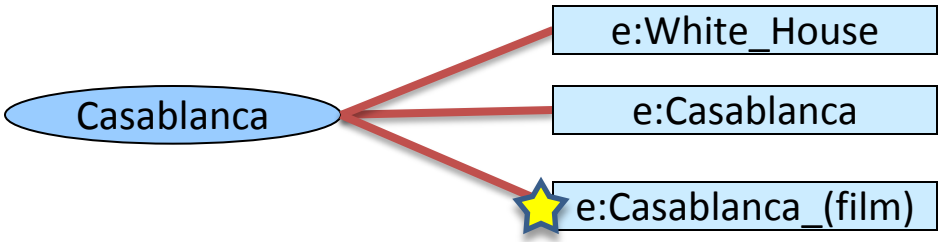


- We rely on human judges

Datasets

- **QALD-1**
  - YAGO2
  - **27/50 questions within scope**
  - *"Which software has been published by Mean Hamster Software?"*
- **NAGA [Elbassuoni et al. CIKM'09]**
  - YAGO+IMDB
  - **44/87 questions within scope**
  - *"Which director has won the Academy Award for Best director and is married to an actress that has won the Academy Award for Best Actress?"*

Results 1/3: Disambiguation



Benchmark	QALD-1	NAGA
cov_{macro}	0.973	0.934
$prec_{macro}$	1.000	0.934
cov_{micro}	0.963	0.945
$prec_{micro}$	1.000	0.941

$$cov = correct/ideal$$
$$prec = correct / retrieved$$

Results 2/3: Query Generation

Issues:

- Incorrect disambiguation
- Incorrect dependencies.

Benchmark	QALD-1	NAGA
cov_{macro}	0.975	0.894
$prec_{macro}$	1.000	0.941
cov_{micro}	0.963	0.847
$prec_{micro}$	1.000	0.906

$$cov = \#correct\ triples / \#ideal\ triples$$
$$prec = \#correct\ triples / \#retrieved\ triples$$

Results 3/3: Answering

Relaxation: Keep type constraints, everything else to keywords

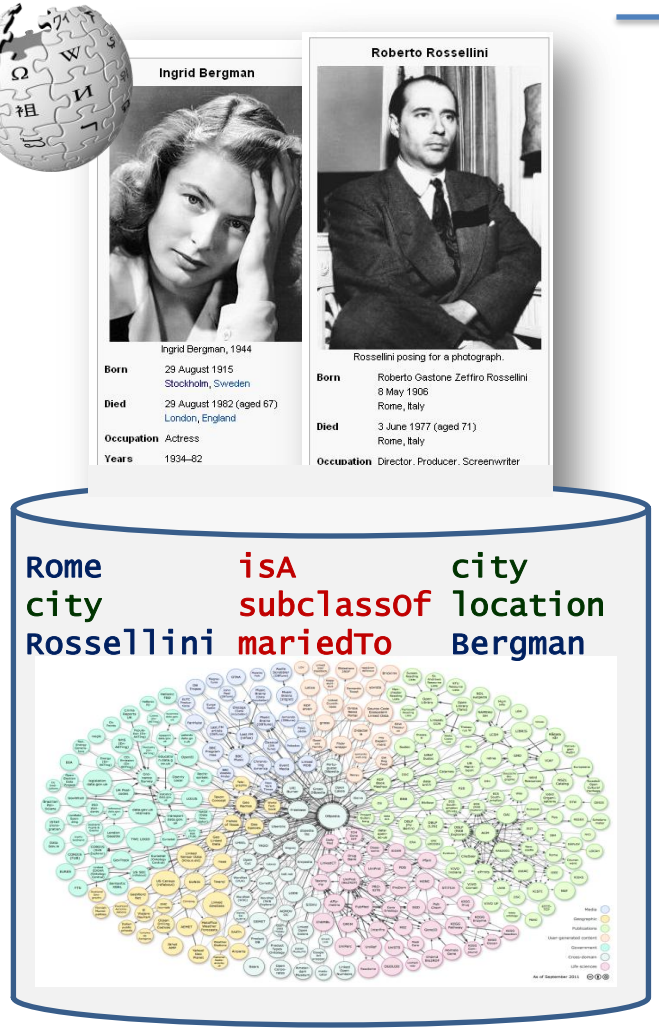
```
?x type writer . ?x bornIn Rome

?x type writer . ?x bornIn ?y ["Rome"]
```

Benchmark	QALD-1	NAGA
#questions	27	44
#queries	20	41
#satisfactory	10	15
#relaxed	+3	+3

unsatisfactory: main problem is **empty result set**

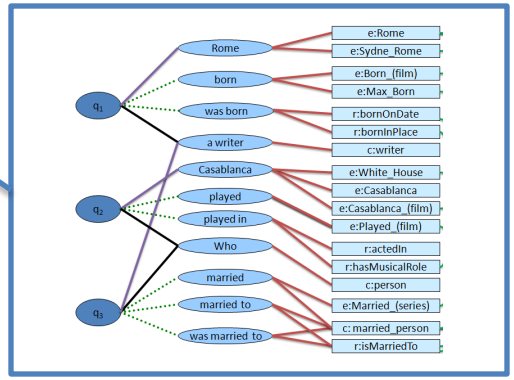
Wrap-up



Question

"Who played in Casablanca and was married to a writer born in Rome?"

DEANNA



SPARQL

```
?p type person.
?p actedIn Casablanca_(film).
?p isMarriedTo ?w.
?w type writer .
?w bornIn Rome
```

KB

Answers

Thank you.

QUESTIONS/COMMENTS?

bit.ly/mpi-deanna