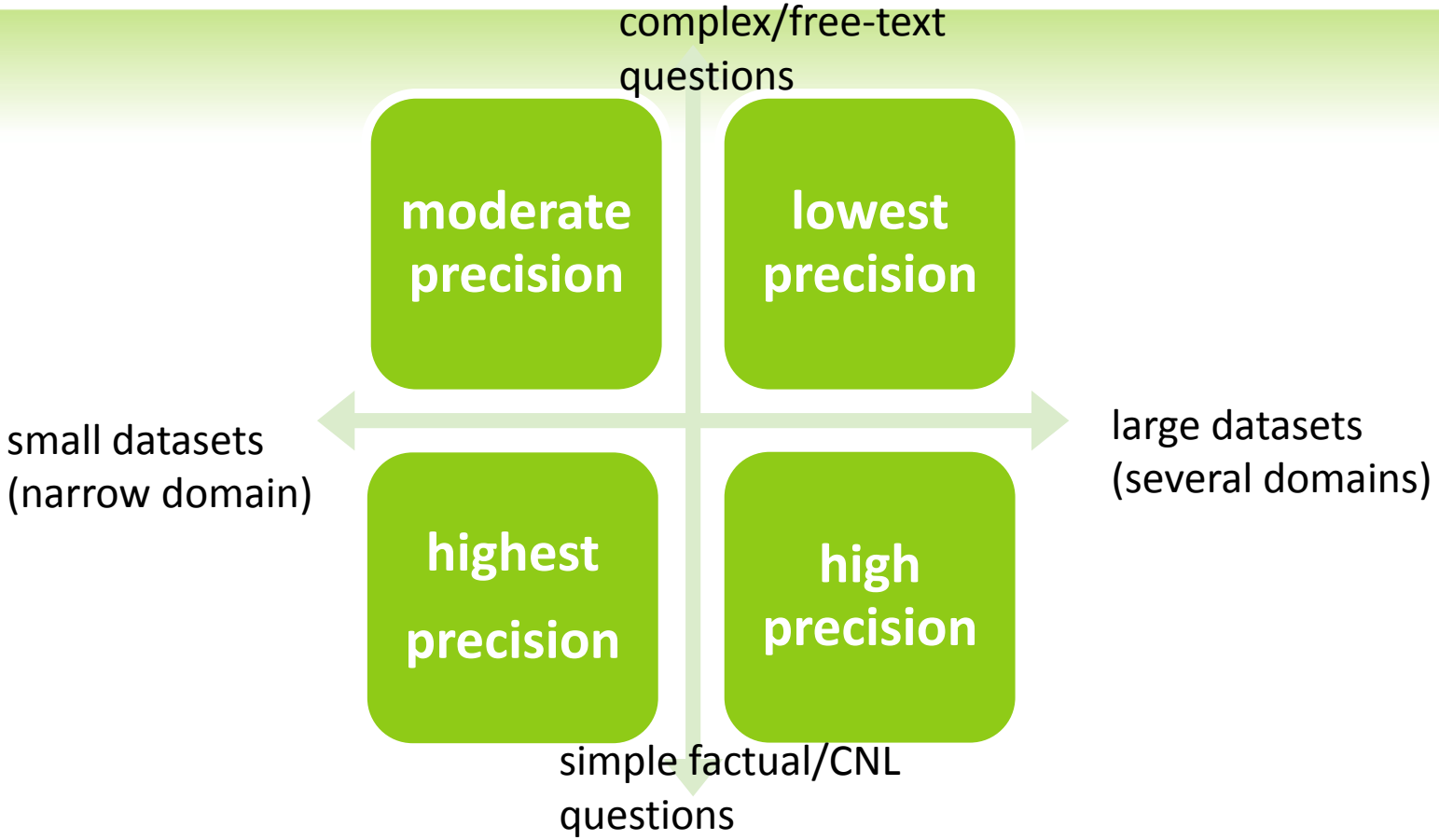
Towards Portable Controlled Natural Languages for Querying Ontologies

Danica Damljanović

email: danica@dcs.shef.ac.uk

Motivation

Portability vs. Performance



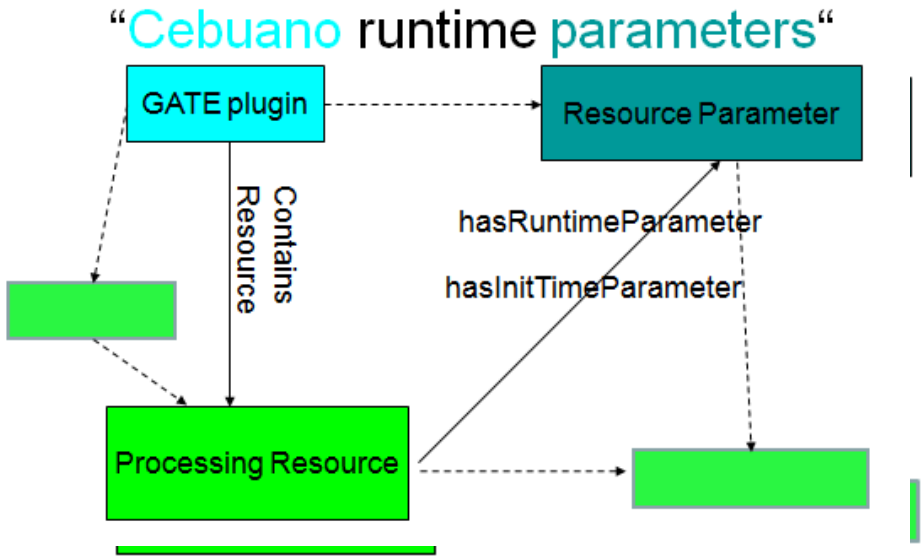
Damljanovic, D., Bontcheva, K.: *Towards Enhanced Usability of Natural Language Interfaces to Knowledge Bases.* In Devedzic V. and Gasevic D. (Eds.), Special issue on Semantic Web and Web 2.0, Annals of Information systems, Springer-Verlag, 2009.

Why portability is a challenge?

- Knowledge representation:
  - using text editor (e.g. by writing OWL)
  - from scratch using ontology editors (Protege),
  - from text (ontology learning tools e.g. Latino),
  - from relational databases,
  - from structured Webpages (Wikipedia Infoboxes),
  - using Controlled Natural Languages (ACE, Rabbit, CLOnE, SOS)

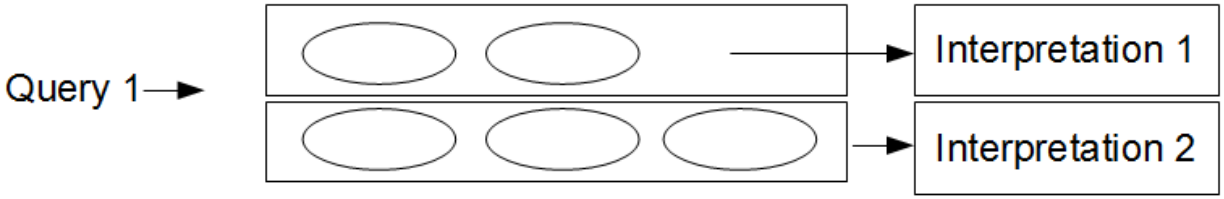
KNOWLEDGE STRUCTURE

- Natural Language vs. formal language

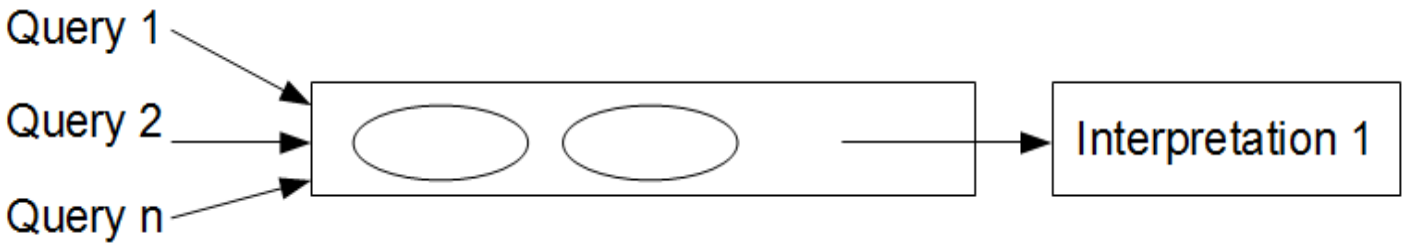


Challenges: Natural Language

Ambiguity



Expressiveness



Natural Language Interfaces for Querying Ontologies

- Controlled Natural Language (CNL): a limited subset of a Natural Language which is translated into a formal language
  - supports certain **vocabulary** and **grammar**
  - is balancing **ambiguity** and **expressiveness**
- **portability**: building the vocabulary (lexicon) automatically from the ontology structure
  - but...

The user's vocabulary

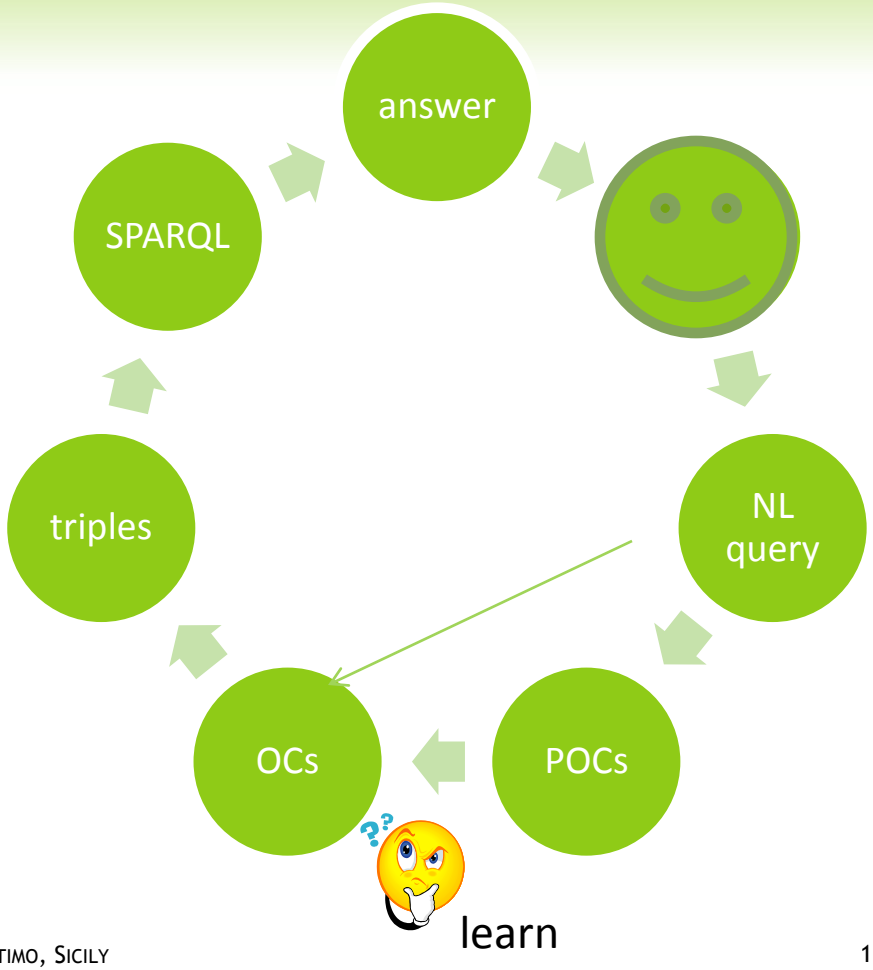
- the lexicalisations in the ontology often do not match those used in the users' questions
  - vocabulary can be extended by using tools such as Wordnet for synonyms
  - still...
    - especially for very specific domains, Wordnet would not find usable lexicalisations - but what about the user's vocabulary?

FREyA (Feedback, Refinement, Extended Vocabulary Aggregator)

- **Feedback**: showing the user system's interpretation of the query
- **Refinement**:
  - resolving ambiguity: generating the dialog whenever one term refers to more than one concept in the ontology (precision)
- **Extended Vocabulary**:
  - expressiveness: generating the dialog whenever an "unknown" term appears in the question (recall)
- The dialog is generated by combining the language of the user and the ontology. Learn from the user's selections.

FREyA Workflow

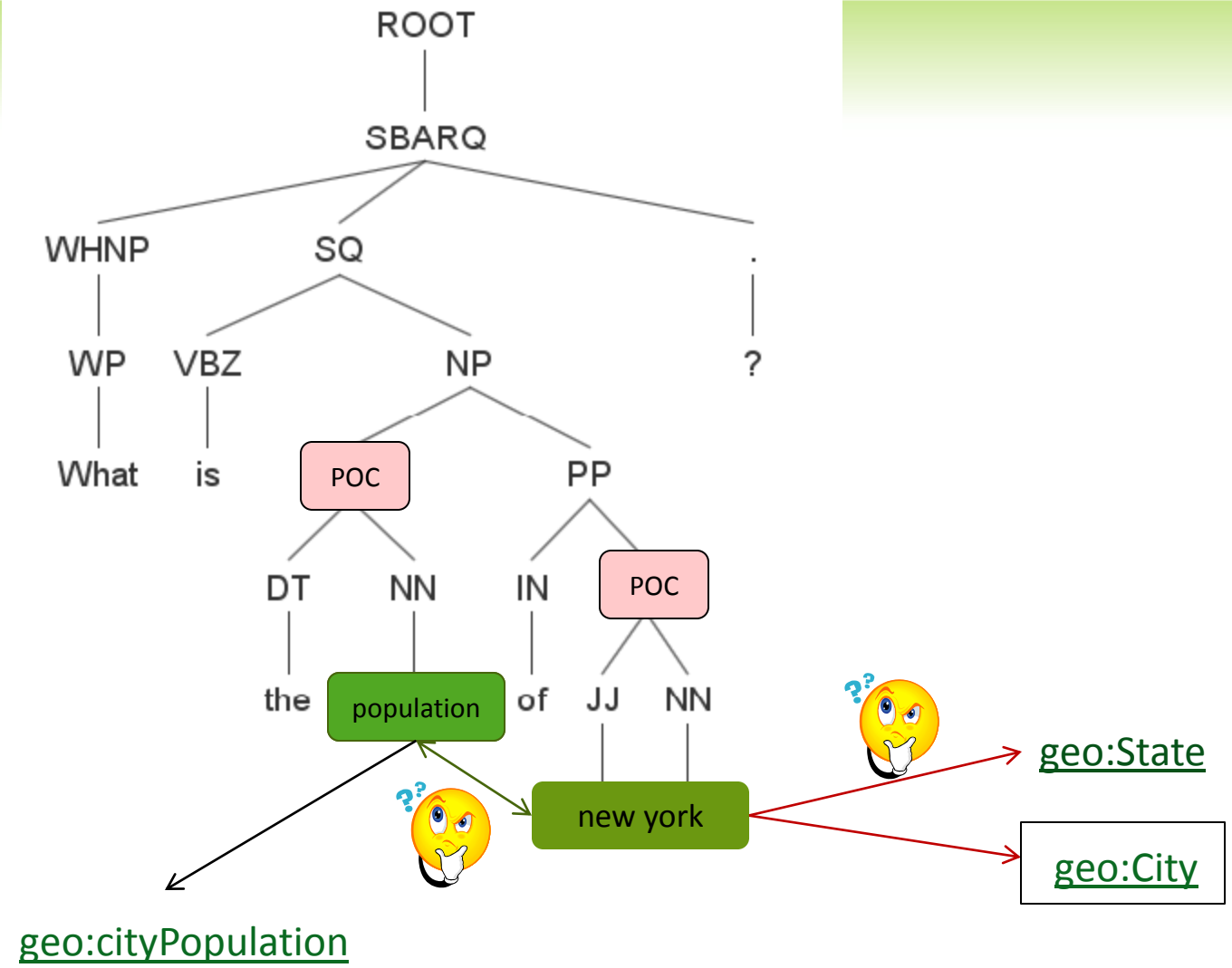
- Potential Ontology Concept (POC)
- Ontology Concept (OC)



Generating Lexicon

- Extract all ontology lexicalisations (lemmas)
- Perform Lexicon-based lookup
- Analyse grammar to find Potential Ontology Concepts (POCs)
- Generate the dialog
- Add the POC to the lexicon

Mapping POC to OCs: Ambiguities



New York is a city

Query: What is the population of new york? Submit

I struggle with new york. Is 'new york' related to:	
city	○
state	○

Query: What is the population of new york? Submit

I struggle with population. Is 'population' related to:	
city population	○
state	○
is city of	○
none	○

New York is a state

Query: What is the population of new york? Submit

I struggle with new york. Is 'new york' related to:	
city	○
state	◉

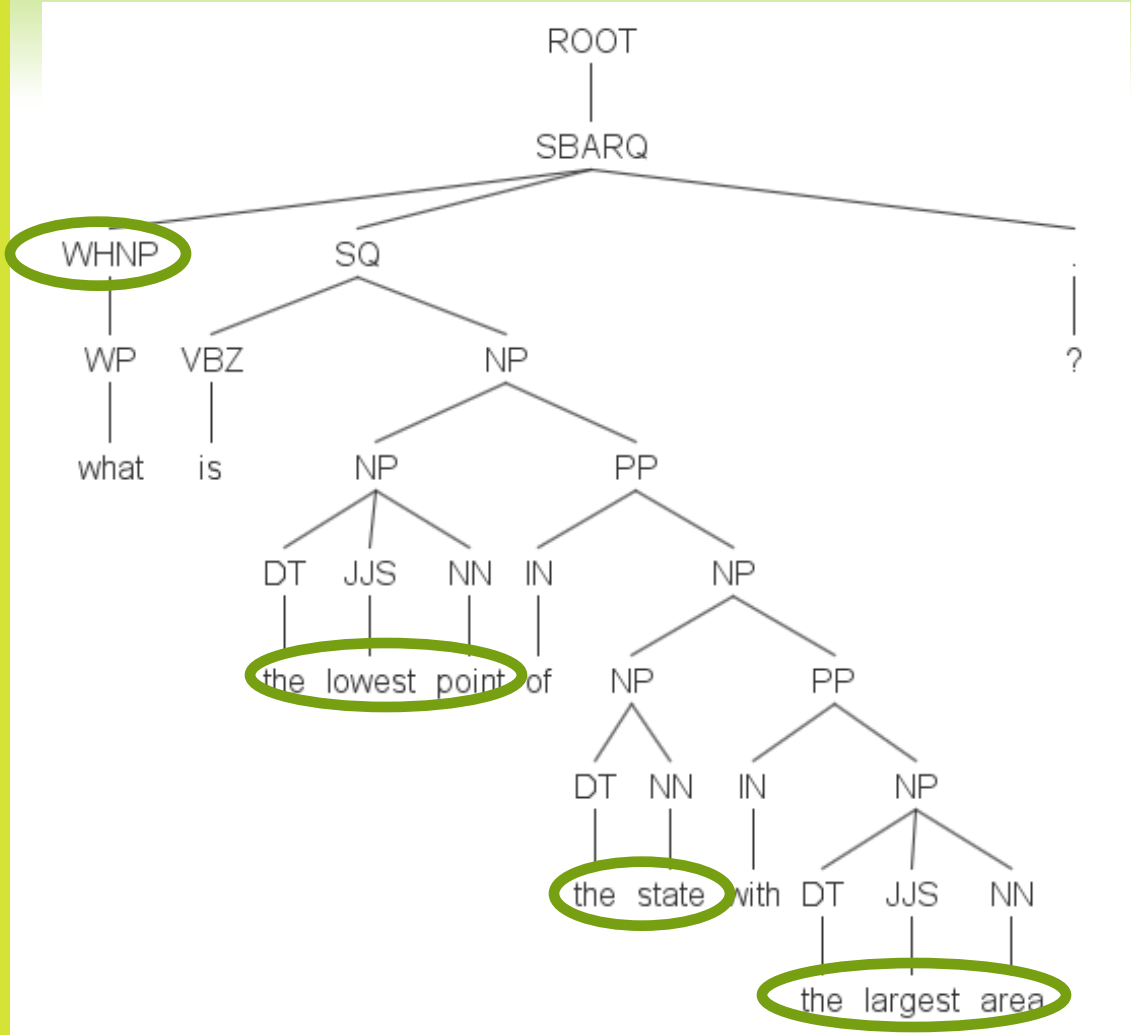
Query: What is the population of new york? Submit

I struggle with population. Is 'population' related to:	
state population	○
state area	○
state pop density	○
none	○

Ambiguous Lexicon

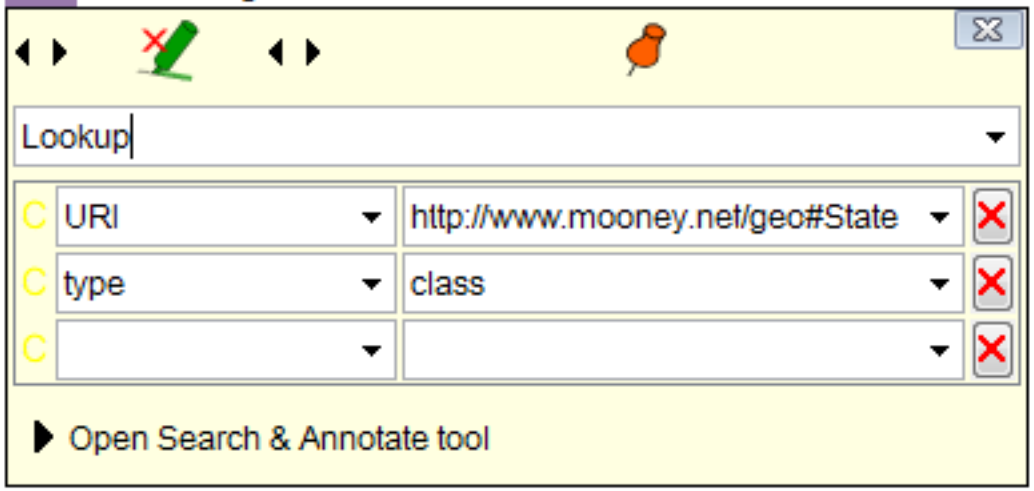
POC	OC (context)	candidate OC	function
new york		geo:State	-
new york		geo:City	-
population	geo:State	geo:statePopulation	-
population	geo:City	geo:cityPopulation	-

Find Potential Ontology Concepts

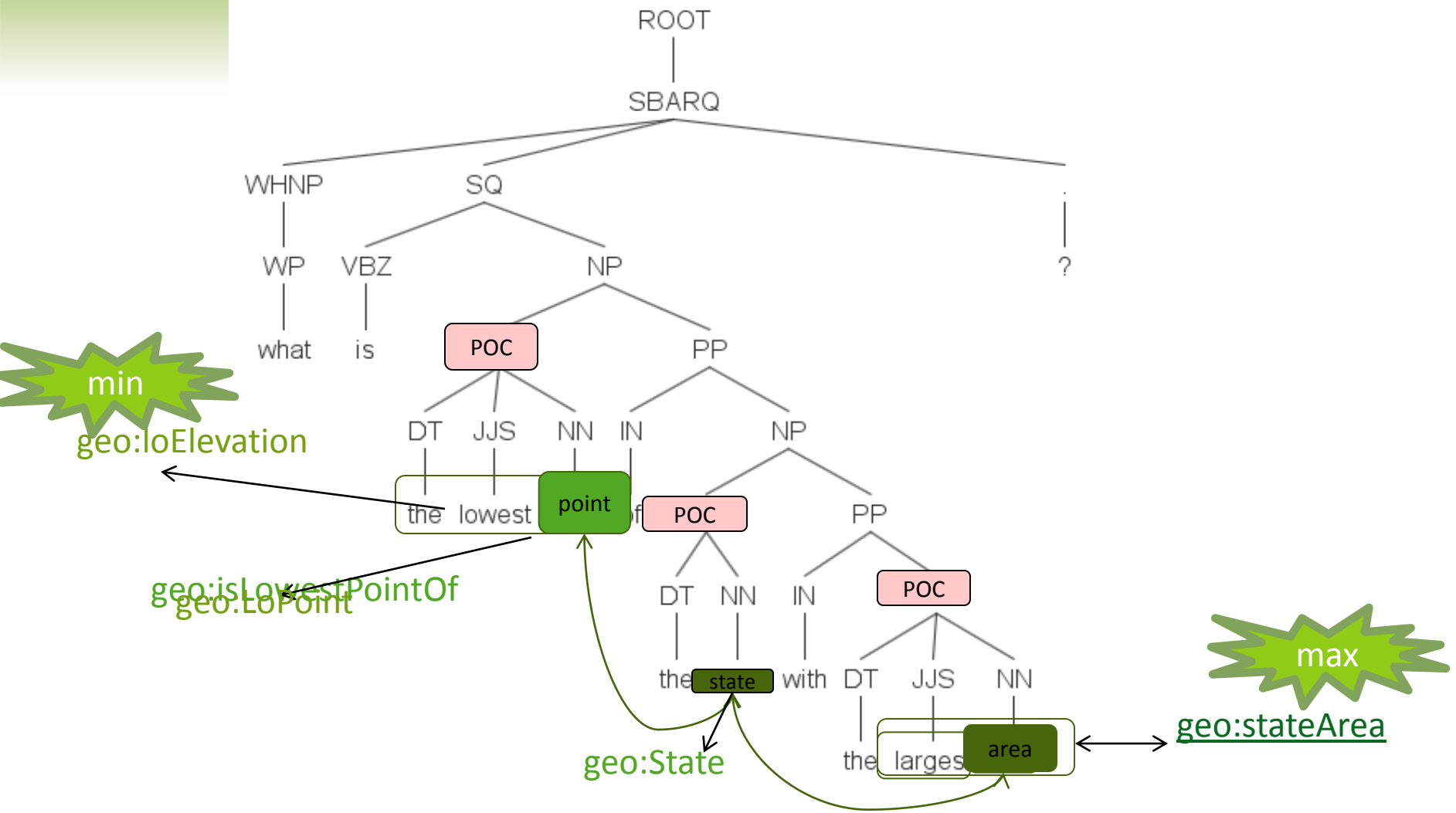


Finding Ontology Concepts

what is the lowest point of the state with the largest area?



The User Controls the Output



What is the lowest point of the state with the largest area?

TRIPLES:

?firstJoker – geo:isLowestPointOf – geo:State

geo:State – (max) geo:stateArea - ?lastJoker

SPARQL:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
select ?firstJoker  ?p0  ?c1  ?p2  ?lastJoker
 where { { { ?c1 ?p0 ?firstJoker}
          UNION { ?firstJoker ?p0 ?c1} .
          filter (?p0=<http://www.mooney.net/geo#isLowestPointOf>) . }
          ?c1 rdf:type <http://www.mooney.net/geo#State> .
          ?c1  ?p2 ?lastJoker .
          filter (?p2=<http://www.mooney.net/geo#stateArea>) . }
 ORDER BY  DESC(xsd:double(?lastJoker))
```

What is the lowest point of the state with the largest area?

the answer for both is Death Valley

TRIPLES:

?firstJoker – (min) geo:loElevation – geo:LoPoint
geo:LoPoint - ?joker3 – geo:State
geo:State – (max) geo:stateArea - ?lastJoker

SPARQL:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
select ?firstJoker  ?p0  ?c1  ?joker3  ?c2  ?p3  ?lastJoker
where { ?c1  ?p0 ?firstJoker .
        filter (?p0=<http://www.moony.net/geo#loElevation>) .
        ?c1 rdf:type <http://www.mooney.net/geo#LoPoint> .
        {{ ?c2  ?joker3 ?c1 }
        UNION  { ?c1  ?joker3 ?c2 }}
       ?c2 rdf:type <http://www.mooney.net/geo#State> .
       ?c2  ?p3 ?lastJoker . filter (?p3=<http://www.mooney.net/geo#stateArea>) . }
ORDER BY  ASC(xsd:double(?firstJoker))  DESC(xsd:double(?lastJoker))
```

New Lexicon

POC	OC (context)	candidate OC	function
area	geo:State	geo:stateArea	-
largest	geo:stateArea	geo:stateArea	max
point	geo:State	geo:LoPoint	-
lowest	geo:LoPoint	geo:loElevation	min
lowest	geo:LoPoint	geo:isLowestPointOf	-

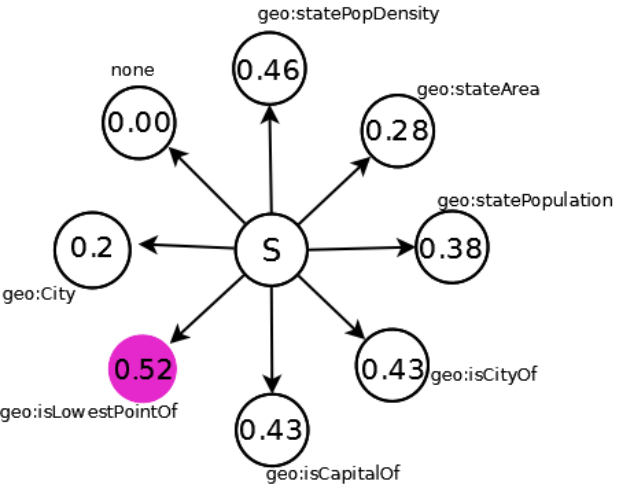
attach scores to each candidate

JSON

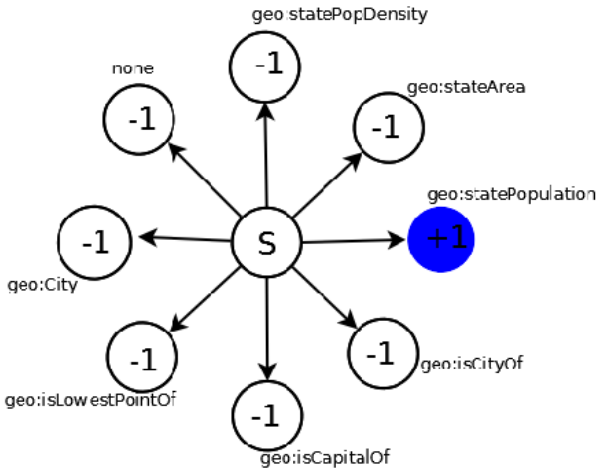
"Key:
area
http://www.mooney.net/geo#State",
"identifier":
"http://www.mooney.net/geo#stateArea",
"function":“ “, “score”:”0.89”

"Key:
largest
http://www.mooney.net/geo#State",
"identifier":
"http://www.mooney.net/geo#stateArea",
"function":“max “, “score”:”0.89”

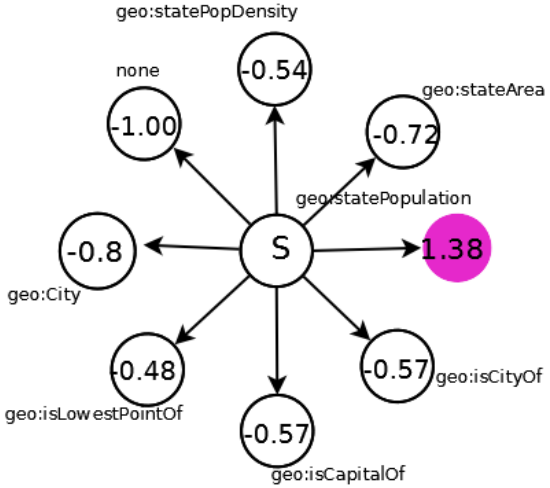
LEARNING



a) INITIAL RANKING



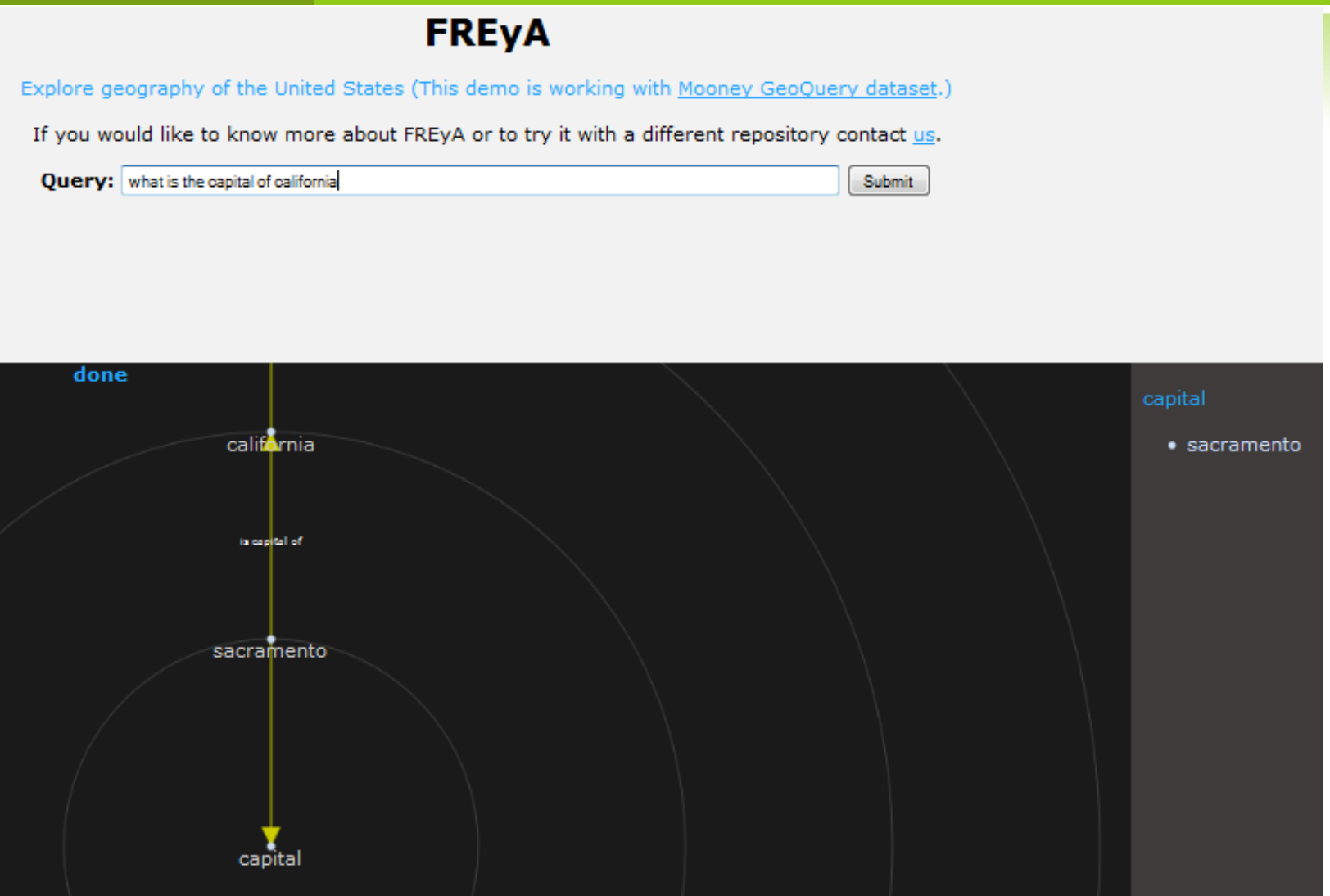
b) REINFORCEMENT BASED ON THE USER SELECTING geo:statePopulation



c) RANKING AFTER THE USER SELECTS geo:statePopulation

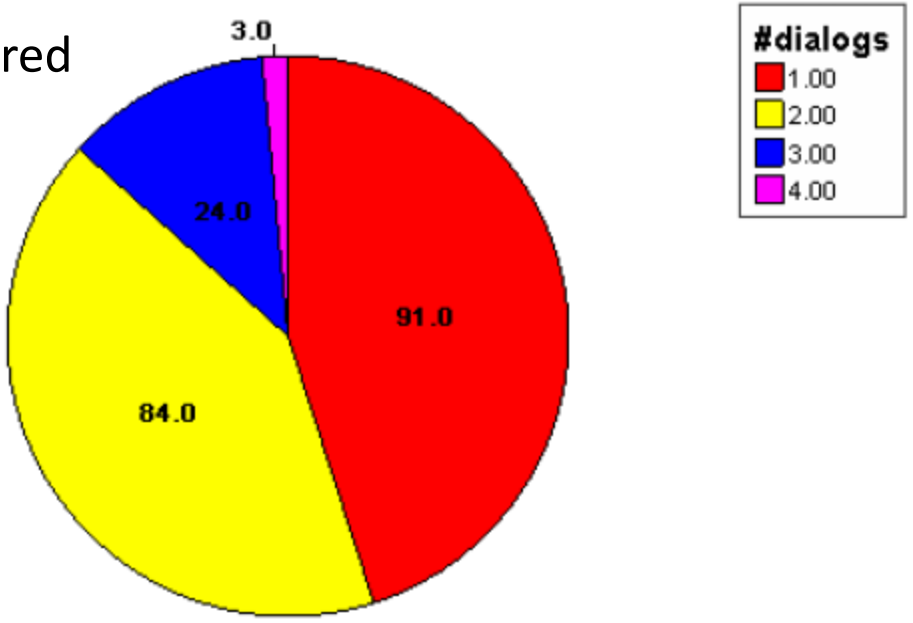
FREyA: a Natural Language Interface to Ontologies

http://gate.ac.uk/freya



EVALUATION: CORRECTNESS

- Mooney GeoQuery dataset, 250 questions
- 34 no dialog, 14 failed to be answered
- Precision=recall=94.4%



Evaluation: Learning

- 10-fold cross-validation with 250 Mooney GeoQuery dataset
- Errors:
  - ambiguity
  - sparseness

Fold	0	1	2	3	4	5	6	7	8	9	Avg
Baseline	.3	.15	.2	.25	.24	.3	.3	.35	.15	.19	0.2476
Learning	.65	.4	.65	.4	.24	.55	.5	.6	.35	.48	0.48

Conclusion

- FREyA combines feedback, refinement and extended vocabulary in order to improve the precision and recall
- the learning model is saved and can be exported/used by any other CNLs for querying ontologies

Next steps

- Improvement of the learning model to avoid errors due to ambiguities
  - point> geo:HiPoint or geo:LoPoint
- Using lexicon to improve other systems

THANK YOU FOR YOUR ATTENTION! QUESTIONS?

Thanks to Abraham Bernstein and Esther Kaufmann from the University of Zurich, for sharing with us Mooney dataset in OWL format, and J. Mooney from University of Texas for making this dataset publicly available.

Contact: danica@dcs.shef.ac.uk

References

- Damljanovic, D., Bontcheva, K.: *Towards Enhanced Usability of Natural Language Interfaces to Knowledge Bases.* In Devedzic V. and Gasevic D. (Eds.), Special issue on Semantic Web and Web 2.0, Annals of Information systems, Springer-Verlag, 2009.