Understanding the tasks of QA over KG

Dennis Diefenbach

The question answering process

| Question analysis | Phrase mapping | Disambiguation | Query construction |

Question analysis | Phrase mapping | Disambiguation | Query construction
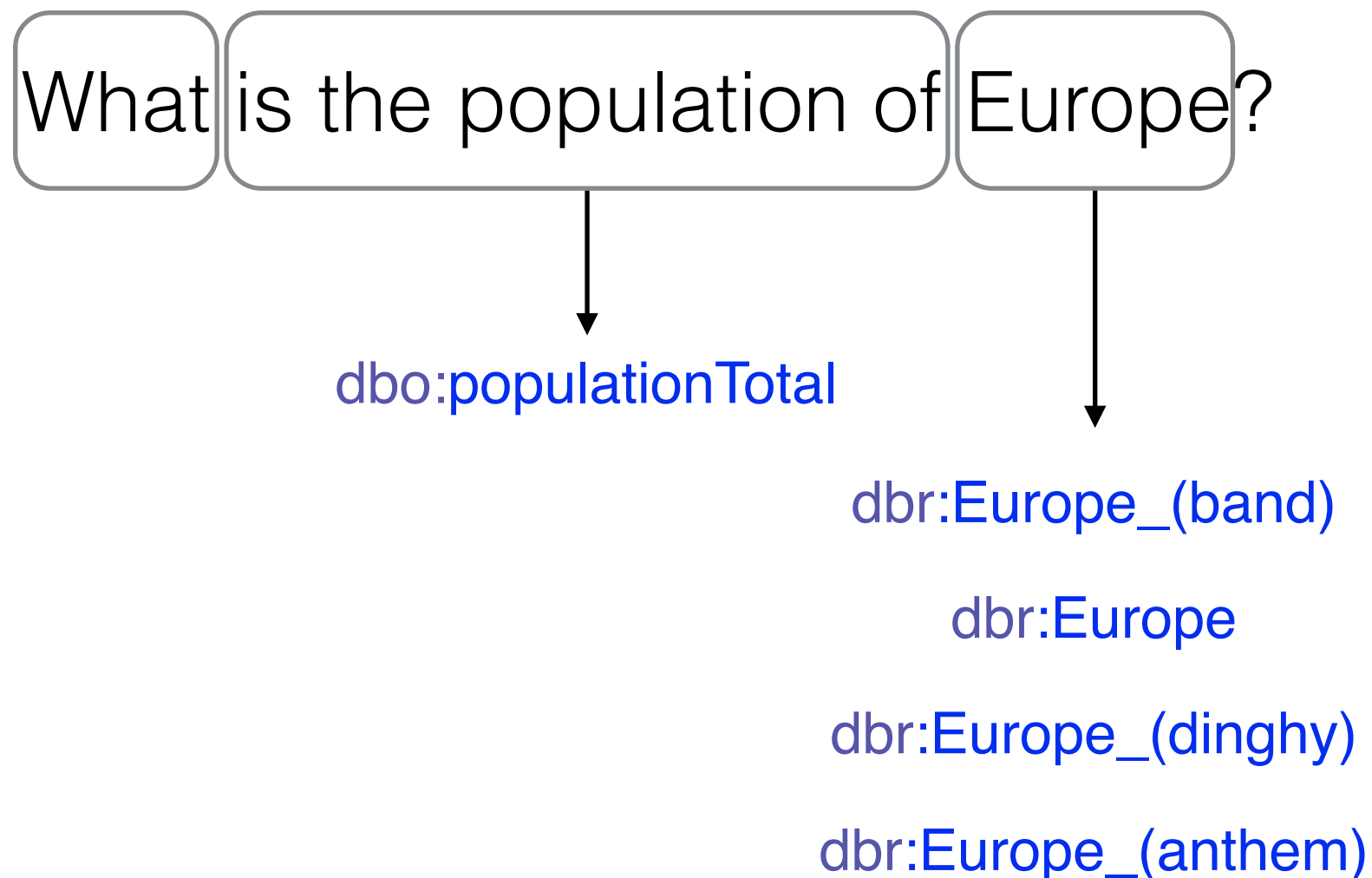
Collect informations which can be deduced considering only the syntax of the question
- Type of the question
- NE recognition
- Identify the properties
- Identify dependencies
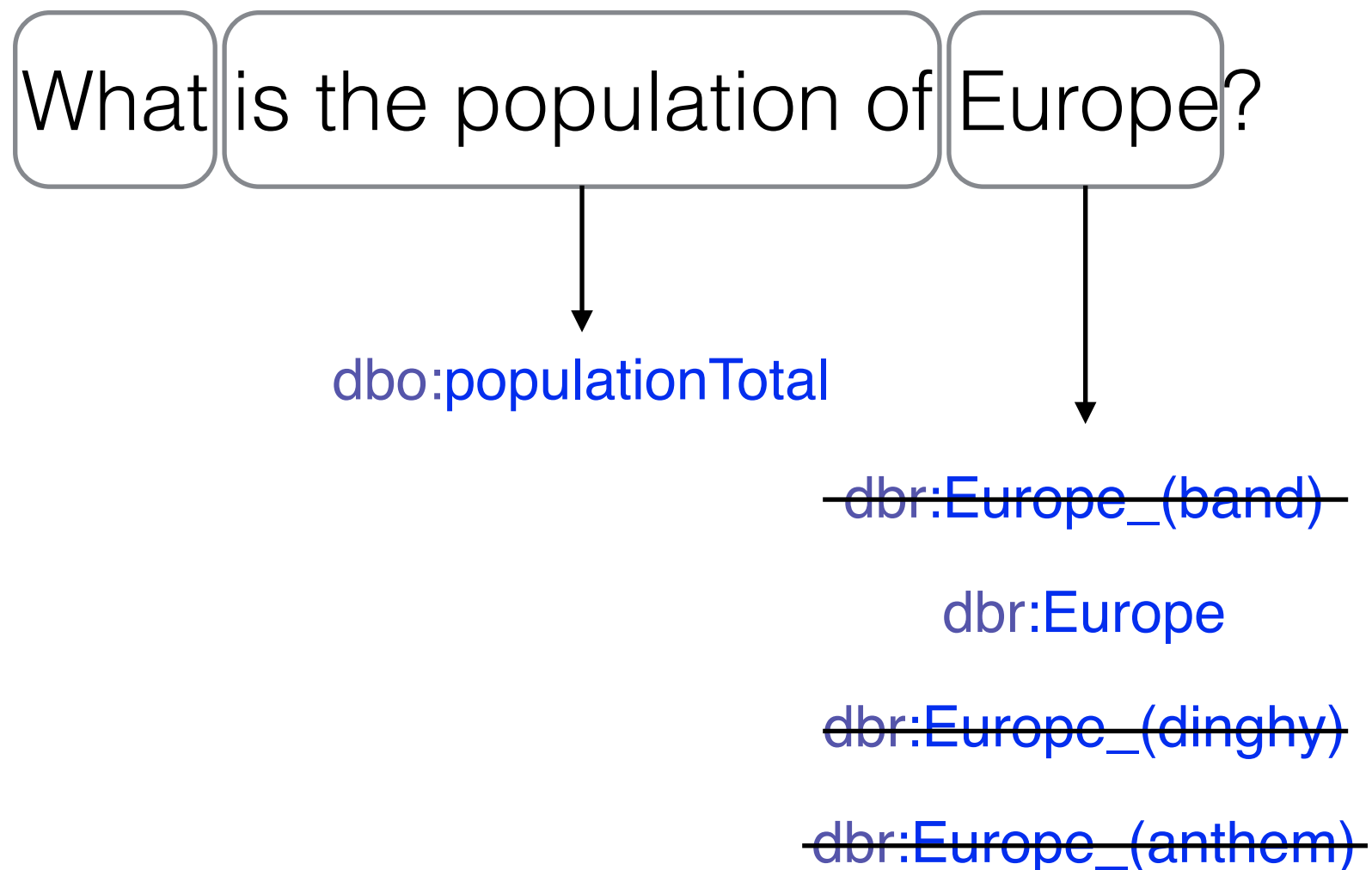
What is the population of Europe?

Question analysis | Phrase mapping | Disambiguation | Query construction

Mapping a phrase to possible resources in the underling ontology

What is the population of Europe?

dbo:populationTotal

dbr:Europe_(band)

dbr:Europe

dbr:Europe_(dinghy)

dbr:Europe_(anthem)

Question analysis · Phrase mapping · **Disambiguation** · Query construction

Mapping a phrase to possible resources in the underling ontology

What | is the population of | Europe?

dbo:populationTotal

~~dbr:Europe_(band)~~

dbr:Europe

~~dbr:Europe_(dinghy)~~

~~dbr:Europe_(anthem)~~

Use all informations collected in the steps before to construct a SPARQL query

What is the population of Europe?

↓

```
Select * where {
    dbr:Europe   dbp:populationTotal   ?p
}
```

NE recognition

Who is the director of the Lord of the Ring?

- Use a NE recognition tool
  - Problem: Standford NER tool could recognize only 51.5% of the NE in the QALD-3 training set
- Check all n-grams
  - Who is the brother of the CEO of the BBC?

use POS Tagging

WRB	VBD	DT	NNP	NNP	VBN	.
When	was	the	European	Union	founded	?

General strategy: identify some reliable POS tags expressions
1. Hand made rules
2. Use ReVerb, based on the following regex
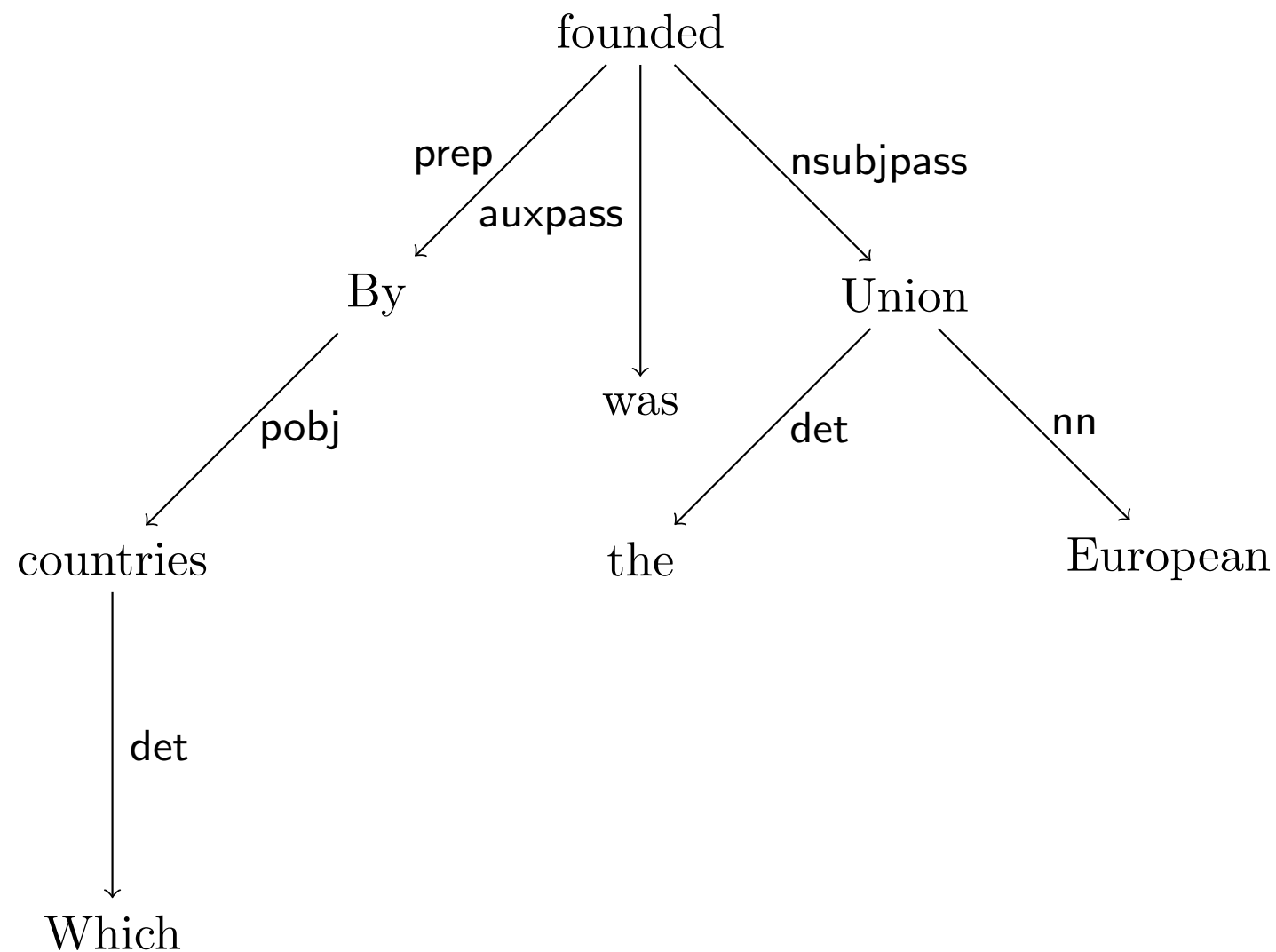
V | VP | VW*P
V = verb particle? adverb?
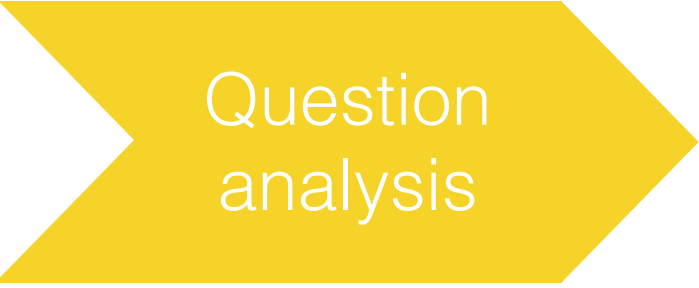W = (noun | adjective | adverb | pronoun | determiner)
P = (preposition | particle | inf. marker)

use Parsers

- Parsers based on dependency grammars
  - Standford dependencies

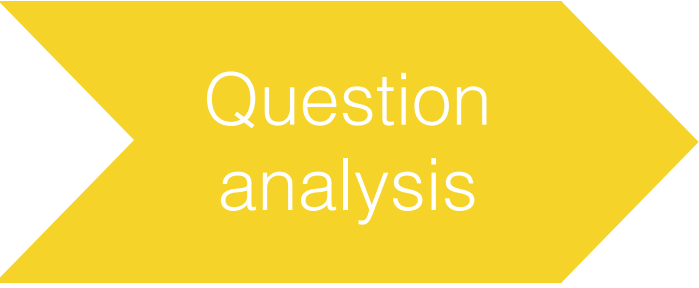# deep neuronal networks

Learn all this from embeddings

Summarizing

Works only for well formulated questions.
Is highly multilingual !!!!

Attention: Which countries are in the European Union?

```
dbr:Greece dbp:member dbr:European_Union .
dbr:France dbp:member dbr:European_Union .
```

For a phrase „s" find, in the underlying KG, a set of resources which correspond to s.

General strategy

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia: <http://dbpedia.org/resource/>;
dbpedia:European_Union    rdfs:label    "European Union"@en
dbpedia:European_Union    rdfs:label    "Europäische Union"@de
dbpedia:European_Union    rdfs:label    "Union européenne"@fr
```

Problems

- Phrase „s" is only similar to the „label(r)"
  - „s" is misspelled
  - order of words in „s" is different
- Phrase „s" is only similar on a semantic point of view to „label(r)"
  - „s" is an abbreviation (e.g. EU)
  - „s" is a nickname (e.g. „Mutti" for „Angela Merkel")
  - „s" is a relational phrase (e.g. „is married with" and „spouse")

Dealing with string similarity

- use Levenstein distance, Jaccrad distance
- use a Lucene Index
  - build in ranking based on tf-idf
  - allows fuzzy searches (searches terms similar to a given metric)
  - hight performant
  - all out of the box

Dealing with semantic similarity

- Database with lexicalizations
  - WordNet, Wiktionary
  - Expand phrase „s" with synonyms (hypernyms/ hyponyms)

Example: EU

{European Union, European Community, EC, European Economic Community, EU, Common Market, Europe}

{europium, Eu, atomic number 63}

Dealing with semantic similarity

- Using large texts
  - wordToVec/ESA
    - Associate to each word a real n-dimensional vector
    - The vector „contains" semantic information!!!
    - ex1. vec(France) near to vec(spain),vec(belgium).
    - ex2. vec(queen) is near to vec(king)-vec(man) +vec(woman)
    - Compare how similar words are by comparing their vectors

Take all triples

What | is the population of | Europe?

dbr:Europe_(band)

dbr:Europe

dbr:Europe_(dinghy)

dbr:Europe_(anthem)

?p   ?o

Templates

What | is the population of | Europe?

Based on the Graph Structure

What | is the population of | Europe?

Benchmarks

Datasets	WebQuestions	SimpleQuestions	QALD 1 to 9
Nb of questions	5.810	108.442	50 to 250
Year of publication	2013	2015	2011 to 2018
Types of relations implied	Reified statements (97%)	Single statements (1 triple)	Up to 3 binary relations
Language	English	English	Multilingual (since 5)
KG	Freebase	Freebase	DBpedia

Benchmarks

Datasets	LC-QuAD	Convex
Nb of questions	5000	5000 dialogs
Year of publication	2017	2019
Types of relations implied	up to 3 triple patterns	?
Language	English	English
KG	DBpedia	Wikidata

Challenges

- Multilinguality

- Portability

- Scalability

- Robustness

- Multiple Knowledge Graphs

- Dialogues

Questions ?