Master Thesis Seminar

Identifying and Linking Counting Quantifiers in Knowledge Bases

Shrestha Ghosh
Department of Computer Science
Saarland University

Advisors
Simon Razniewski
Gerhard Weikum

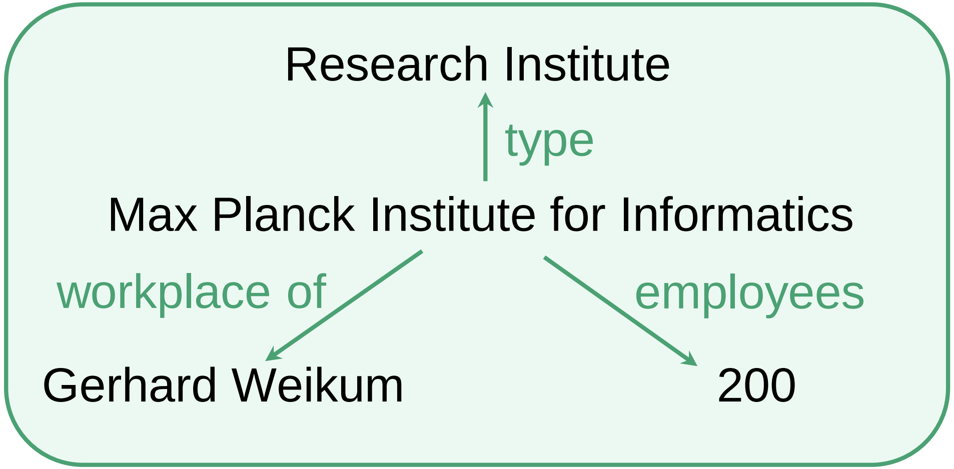
Outline

1. Introduction
2. Motivation
3. Problem Statement
4. Research Questions
5. Challenges
6. Related Work
7. KB Selection
8. Identification and Extraction
9. Predicate Alignment
10. Alignment Evaluation
11. Extension
12. Summary

1. Introduction

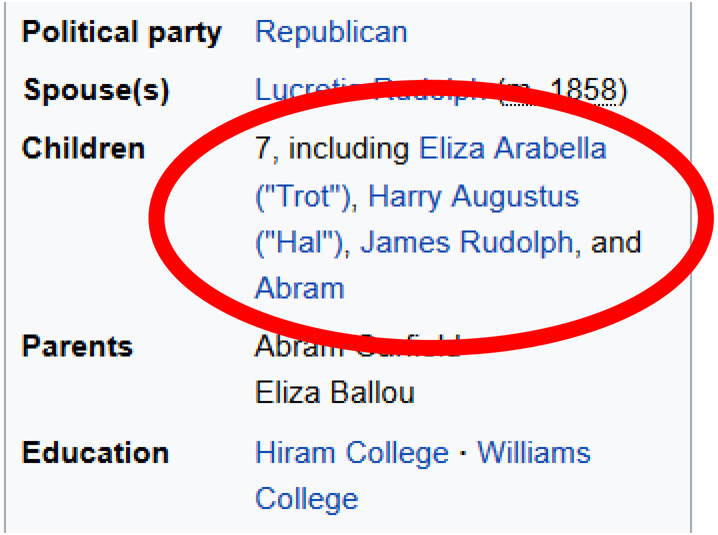
Knowledge Bases

- Provide structured information on items



2. Motivation 1. KBs mix count and standard facts

Snippet from the Wikipedia page of James A. Garfield

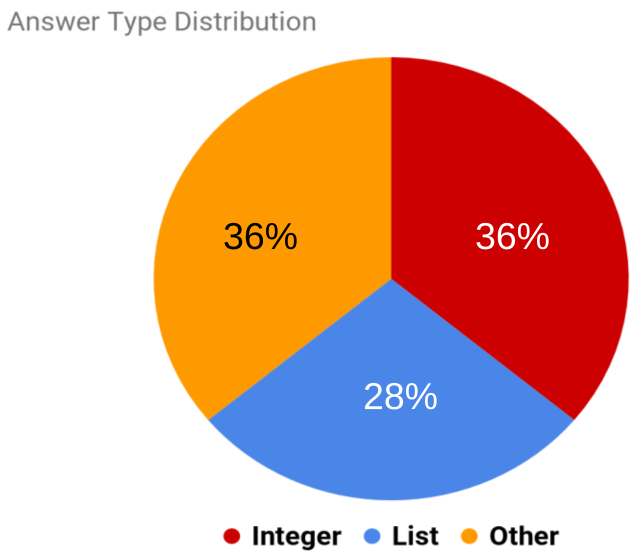


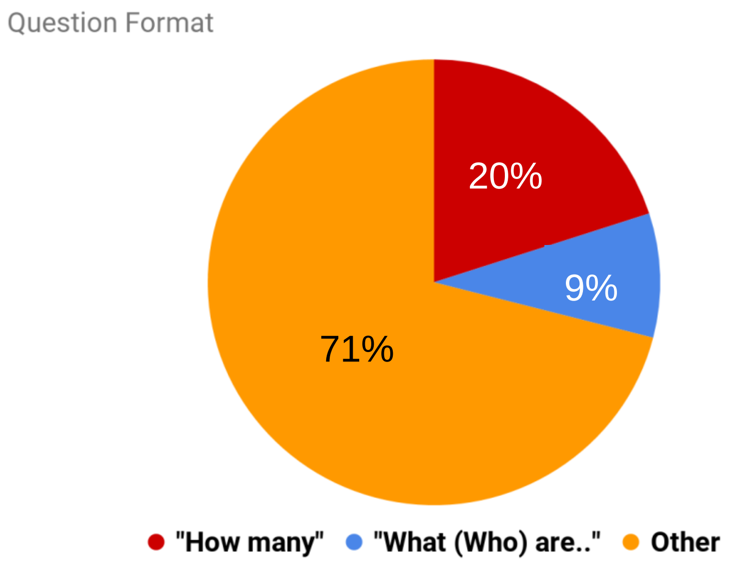
Political party	Republican
Spouse(s)	Lucretia Rudolph (m. 1858)
Children	7, including Eliza Arabella ("Trot"), Harry Augustus ("Hal"), James Rudolph, and Abram
Parents	Abram Garfield Eliza Ballou
Education	Hiram College · Williams College

Out of 70k values for the children predicate in DBpedia, 33% are integers

2. Motivation 2: Questions frequently concern counts

Free917 Dataset contains: 641 train and 276 test Q&A pairs





2. Motivation 3: Count questions can be enriched by facts (and vice versa)

Q: "Employees at Max Planck Institute for Informatics"

A: 200

Q: Who are these?

A: Not exactly modelled as *‘employed by’* or *‘employed at’*, but, related predicates:

- workplace of: Gerhard Weikum, ...
- director: Kurt Mehlhorn, Bernt Schiele, ...

3. Problem Statement

- Direction: integer valued ➔ entity valued predicates (previous example of employees)
  - Employees ➔ workplace of, director
- Direction: entity valued ➔ integer valued predicates
  - Example Q.: What are the moons of Jupiter?
  - Standard QA answer: “Ganymede, Callisto, Europa, Io, ...”
  - Enhancing with related predicate “number of moons”:

    “79, some of which are Ganymede, Io, ..”

Goal

Investigate implicit count information by identifying entity valued predicates and integer valued predicates and aligning semantically related pairs.

4. Research Questions

1. How to identify and extract counting information from structured sources?

   children ✓ employees ✓ workplace ✓ interests ✗ pageCount ✗

2. How to align related entity and integer valued predicates?

   children ↔ child of^{-1}, parent of, child

   employees ↔ workplace, staff at

3. How to evaluate the alignment?

5. Challenges

1. Alignment itself
   a. Semantic and statistic alignment
   b. Fuzzy alignment
2. Aligning dirty data
   a. Predicates which do not have a clear preference of object types, e.g., children
3. Dealing with inverse predicates, their classification
4. Dealing with unknown
   a. Identify comma separated words, city names and person names from strings

6. Related Work

1. Cardinality from text sources - Mirza et al. [1]ACL 2017, [2]ISWC 2018
2. Cardinality scores for assessing KB completeness - Tanon et al.[3], ISWC 2017
3. Numerical Open IE - Saha et al.[4], ACL 2017
4. Ontology alignment - Rahm et al.[5], VLDB 2001
5. Current QA systems treatment of count information - Bast et al.[6], ACM 2015

Mirza et al. [1] "Cardinal Virtues: Extracting Relation Cardinalities from Text.", [2] “Enriching Knowledge Bases with Counting Quantifiers”;
[3] Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs.”; [4] Saha et al. “Bootstrapping for Numerical Open IE.”;
[5] Rahm et al. “A Survey of Approaches to Automatic Schema Matching ”; [6] Bast et al. "More accurate question answering on freebase."

6. Related Work

1. Cardinality from text sources - Mirza et al. [1]ACL 2017, [2]ISWC 2018
2. Cardinality scores for assessing KB completeness - Tanon et al.[3], ISWC 2017
3. Numerical Open IE - Saha et al.[4], ACL 2017
4. Ontology alignment - Rahm et al.[5], VLDB 2001
5. Current QA systems treatment of count information - Bast et al.[6], ACM 2015

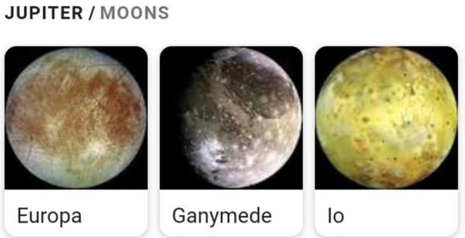
What are the moons of Jupiter?

Mirza et al. [1] "Cardinal Virtues: Extracting Relation Cardinalities from Text.", [2] “Enriching Knowledge Bases with Counting Quantifiers”;
[3] Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs.”; [4] Saha et al. “Bootstrapping for Numerical Open IE.”;
[5] Rahm et al. “A Survey of Approaches to Automatic Schema Matching ”; [6] Bast et al. "More accurate question answering on freebase."

6. Related Work

1. Cardinality from text sources - Mirza et al. [1]ACL 2017, [2]ISWC 2018
2. Cardinality scores for assessing KB completeness - Tanon et al.[3], ISWC 2017
3. Numerical Open IE - Saha et al.[4], ACL 2017
4. Ontology alignment - Rahm et al.[5], VLDB 2001
5. Current QA systems treatment of count information - Bast et al.[6], ACM 2015

What are the moons of Jupiter?



...

(displays 51 of known 79)

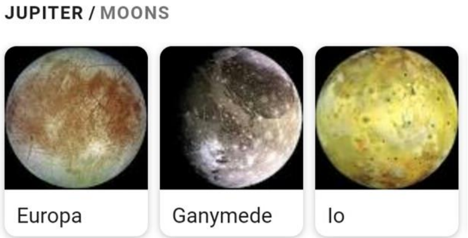
Mirza et al. [1] "Cardinal Virtues: Extracting Relation Cardinalities from Text.", [2] “Enriching Knowledge Bases with Counting Quantifiers”; [3] Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs.”; [4] Saha et al. “Bootstrapping for Numerical Open IE.”; [5] Rahm et al. “A Survey of Approaches to Automatic Schema Matching ”; [6] Bast et al. "More accurate question answering on freebase."

6. Related Work

1. Cardinality from text sources - Mirza et al. [1]ACL 2017, [2]ISWC 2018
2. Cardinality scores for assessing KB completeness - Tanon et al.[3], ISWC 2017
3. Numerical Open IE - Saha et al.[4], ACL 2017
4. Ontology alignment - Rahm et al.[5], VLDB 2001
5. Current QA systems treatment of count information - Bast et al.[6], ACM 2015

What are the moons of Jupiter?

How many employees does Google have?



...

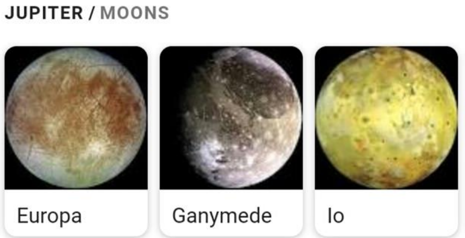
(displays 51 of known 79)

Mirza et al. [1] "Cardinal Virtues: Extracting Relation Cardinalities from Text.", [2] “Enriching Knowledge Bases with Counting Quantifiers”; [3] Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs.”; [4] Saha et al. “Bootstrapping for Numerical Open IE.”; [5] Rahm et al. “A Survey of Approaches to Automatic Schema Matching ”; [6] Bast et al. "More accurate question answering on freebase."

6. Related Work

1. Cardinality from text sources - Mirza et al. [1]ACL 2017, [2]ISWC 2018
2. Cardinality scores for assessing KB completeness - Tanon et al.[3], ISWC 2017
3. Numerical Open IE - Saha et al.[4], ACL 2017
4. Ontology alignment - Rahm et al.[5], VLDB 2001
5. Current QA systems treatment of count information - Bast et al.[6], ACM 2015

What are the moons of Jupiter?



JUPITER / MOONS

Europa

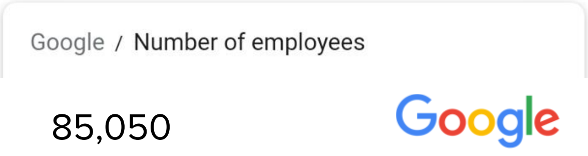
Ganymede

Io

...

(displays 51 of known 79)

How many employees does Google have?



Google / Number of employees

85,050

Mirza et al. [1] "Cardinal Virtues: Extracting Relation Cardinalities from Text.", [2] “Enriching Knowledge Bases with Counting Quantifiers”;
[3] Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs.”; [4] Saha et al. “Bootstrapping for Numerical Open IE.”;
[5] Rahm et al. “A Survey of Approaches to Automatic Schema Matching ”; [6] Bast et al. "More accurate question answering on freebase."

7. KB Selection

- **Wikipedia infoboxes**

  Considerable effort in processing

- **DBpedia raw extraction**

  Cleaner, filters mixed infobox entries down to integer OR entities

- **DBpedia ontology**

  Canonicalised predicates with type constraints

- **Wikidata**

  Very clean with fewer predicates

7. KB Selection

- **Wikipedia infoboxes**

  Considerable effort in processing

- **DBpedia raw extraction**

  Cleaner, filters mixed infobox entries down to integer OR entities

- **DBpedia ontology**

  Canonicalised predicates with type constraints

- **Wikidata**

  Very clean with fewer predicates

7. KB Selection

- **Wikipedia infoboxes**

  Considerable effort in processing

- **DBpedia raw extraction**

  Cleaner, filters mixed infobox entries down to integer OR entities

- **DBpedia ontology**

  Canonicalised predicates with type constraints

- **Wikidata**

  Very clean with fewer predicates

8. Identification and Extraction - Step 1

Source: DBpedia raw extraction + ontology

Question:

Which predicates are entity valued, which ones are integer valued?

Approach:

Look at statistical distribution of datatypes that objects of each predicates take

- Integer, float, named entity, etc.

8. Identification and Extraction - Step 1

Distribution of predicates in DBpedia infobox:

- 60k distinct predicates, 4061 with at least 1k occurrences
- Datatypes that objects take in a spo triple
  - named entity (NE), integer, float, date, comma-separated, unknown
- Clean predicates
  - Predicates whose objects predominantly take one datatype
- Mixed predicates
  - More than one dominant object datatype

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.

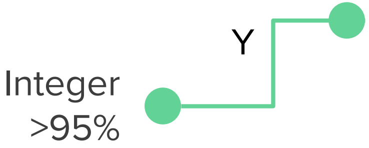
8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.

Integer
>95% ●

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.



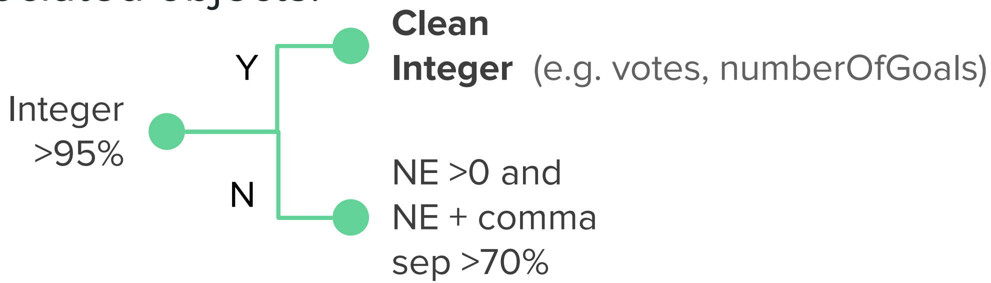
Integer >95%

Y

Clean Integer (e.g. votes, numberOfGoals)

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.



Integer >95%

Y

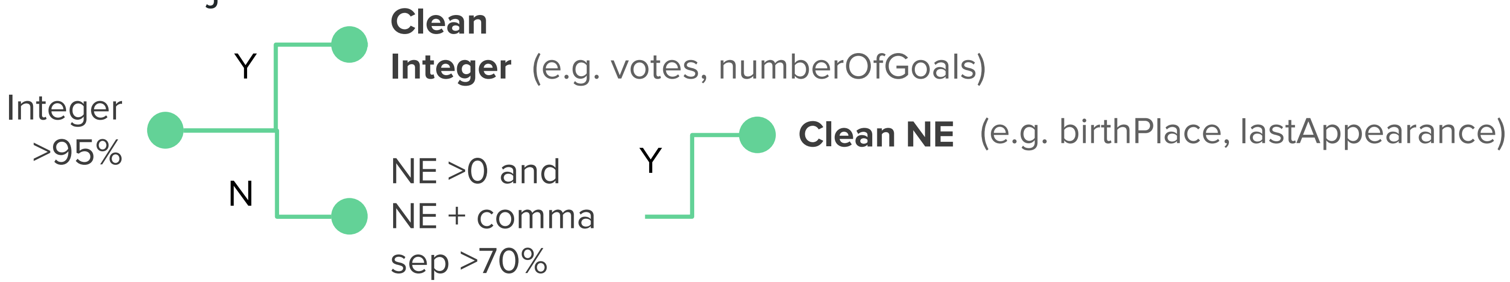
Clean Integer (e.g. votes, numberOfGoals)

N

NE >0 and NE + comma sep >70%

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.



Integer >95%

Y

Clean Integer (e.g. votes, numberOfGoals)

N

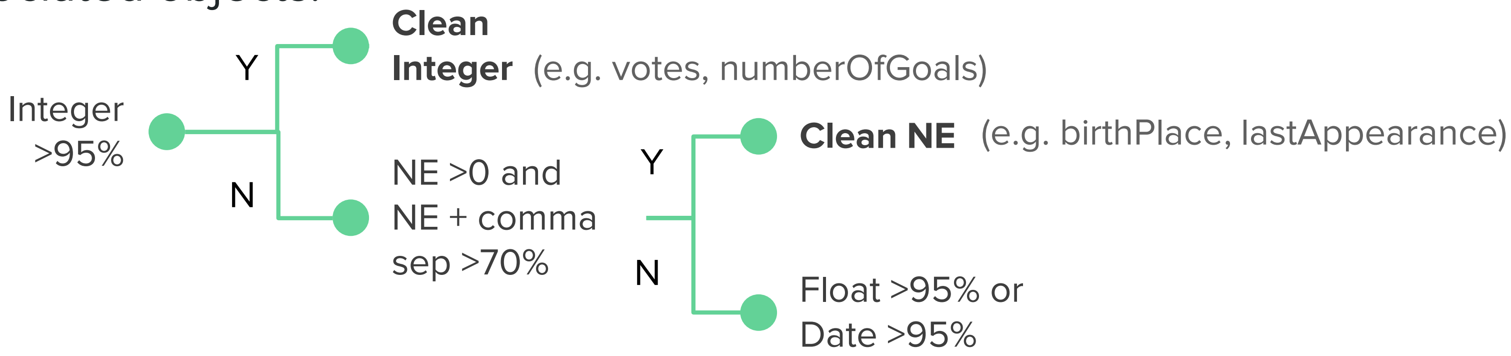
NE >0 and NE + comma sep >70%

Y

Clean NE (e.g. birthPlace, lastAppearance)

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate's associated objects.



Integer >95%

Y

Clean Integer (e.g. votes, numberOfGoals)

N

NE >0 and NE + comma sep >70%

Y

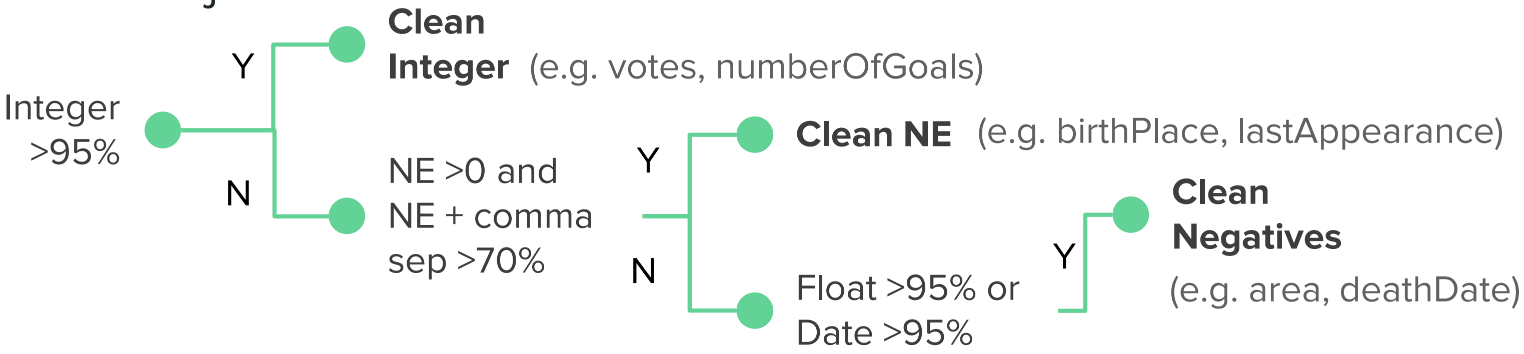
Clean NE (e.g. birthPlace, lastAppearance)

N

Float >95% or Date >95%

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate's associated objects.



Integer >95%

Y

Clean Integer (e.g. votes, numberOfGoals)

N

NE >0 and NE + comma sep >70%

Y

Clean NE (e.g. birthPlace, lastAppearance)

N

Float >95% or Date >95%

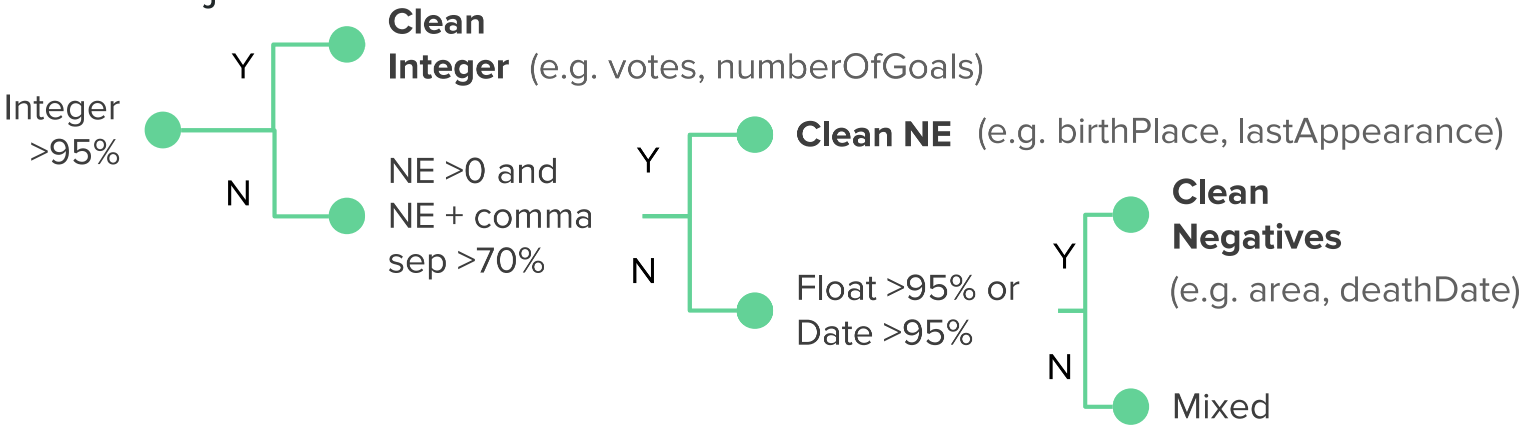
Y

Clean Negatives

(e.g. area, deathDate)

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate's associated objects.



Integer >95%

Y

Clean Integer (e.g. votes, numberOfGoals)

N

NE >0 and NE + comma sep >70%

Y

Clean NE (e.g. birthPlace, lastAppearance)

N

Float >95% or Date >95%

Y

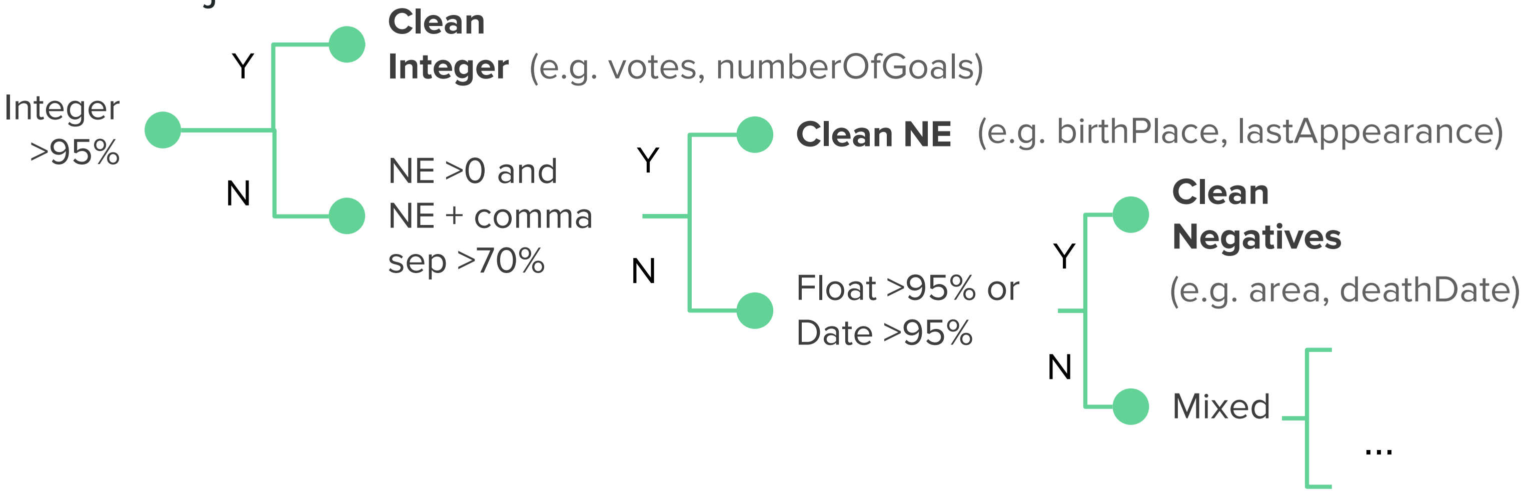
Clean Negatives (e.g. area, deathDate)

N

Mixed

8. Identification and Extraction - Step 1

Clean case classification based on distribution of datatypes of a predicate’s associated objects.


Integer >95%
Y
N
Clean Integer (e.g. votes, numberOfGoals)
NE >0 and NE + comma sep >70%
Y
N
Clean NE (e.g. birthPlace, lastAppearance)
Float >95% or Date >95%
Y
N
Clean Negatives (e.g. area, deathDate)
Mixed
...

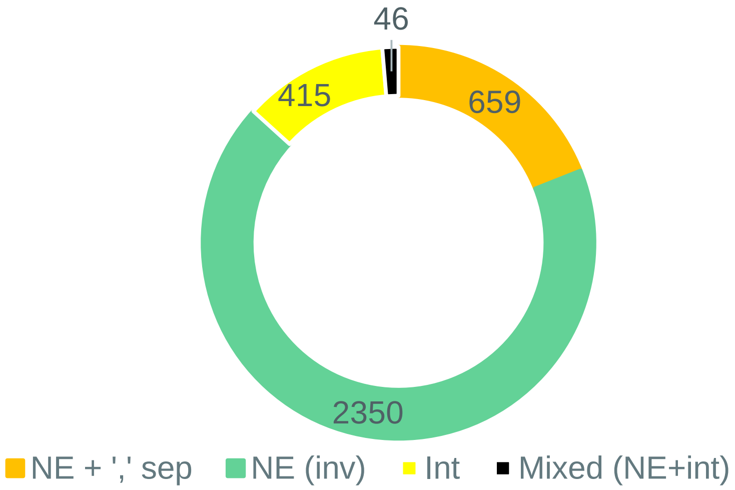
Mixed case classification done similarly into -

Exclusive mix of Int and NE, Dirty mix of Int and NE, Dirty Int, Dirty NE

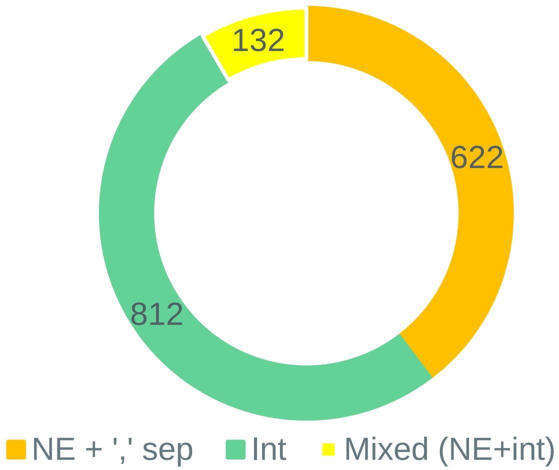
8. Identification and Extraction - Results

Remove predicates which store measurement information or have names comprising less than four letters

a. Latm, longm, height, rank, speed, r1c, ne



Clean Predicates

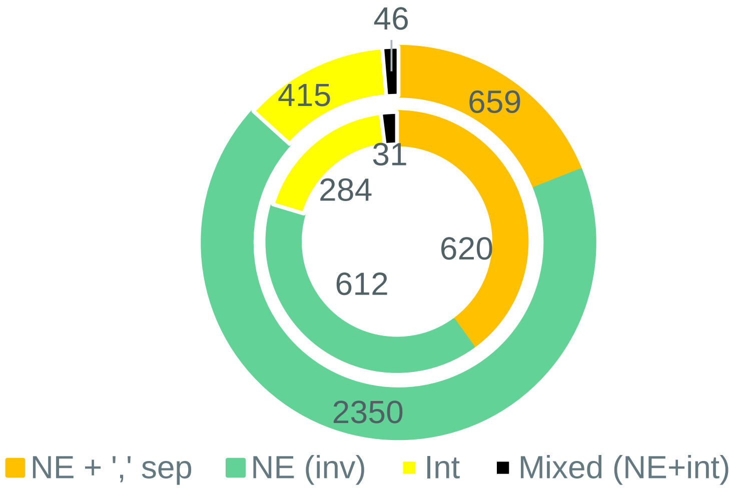


Dirty Predicates

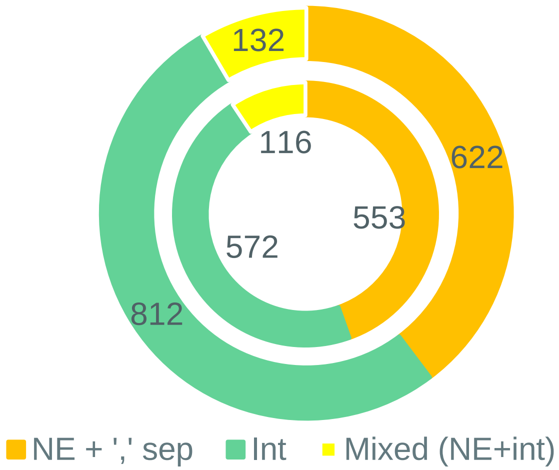
8. Identification and Extraction - Results

Remove predicates which store measurement information or have names comprising less than four letters

a. Latm, longm, height, rank, speed, r1c, ne



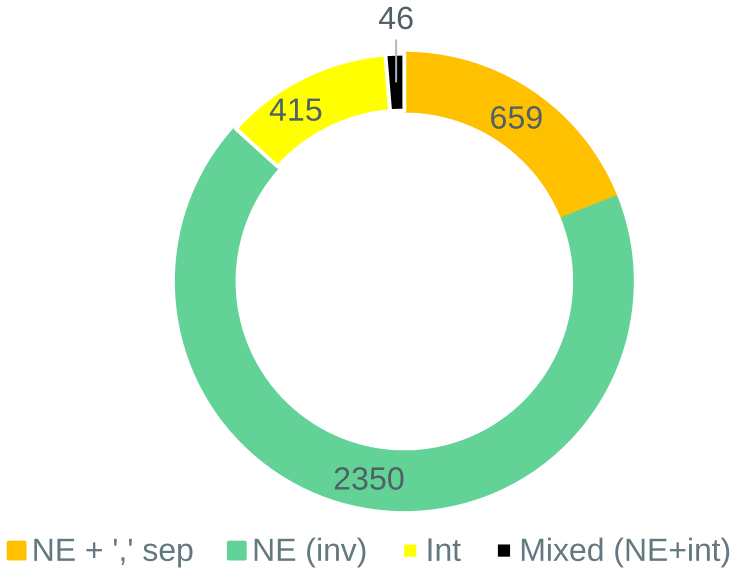
Clean Predicates



Dirty Predicates

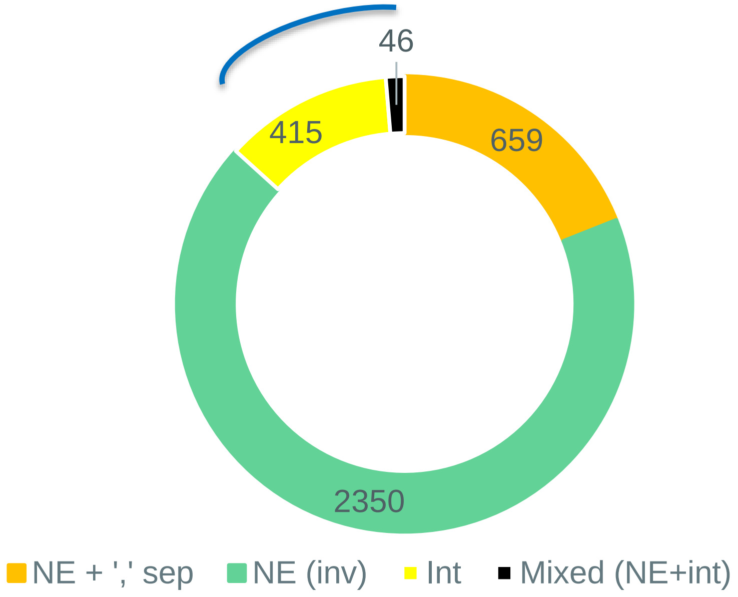
8. Identification and Extraction - Results

Potential candidates for count related predicates



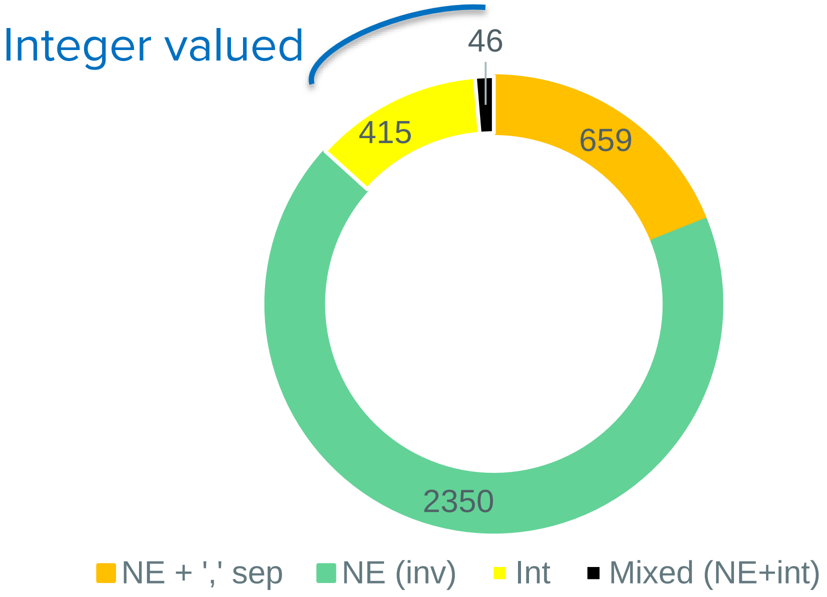
8. Identification and Extraction - Results

Potential candidates for count related predicates



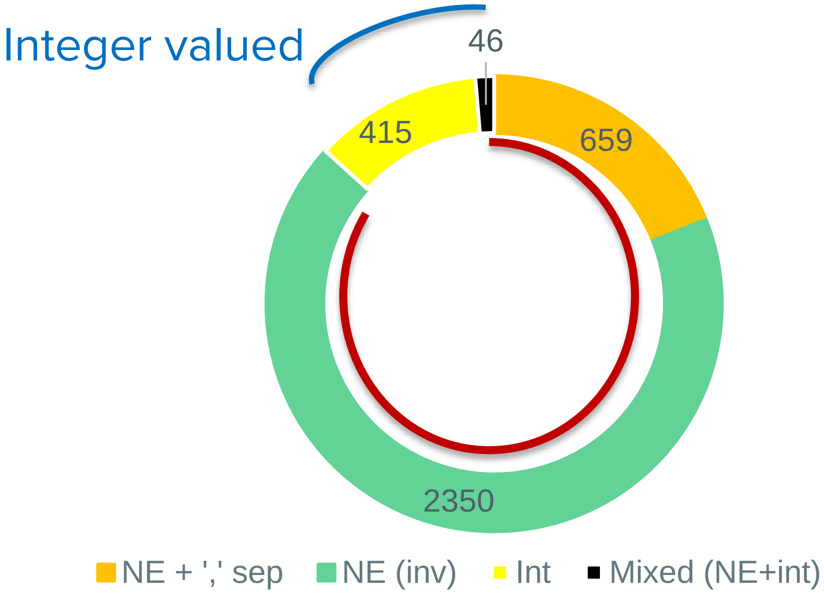
8. Identification and Extraction - Results

Potential candidates for count related predicates



8. Identification and Extraction - Results

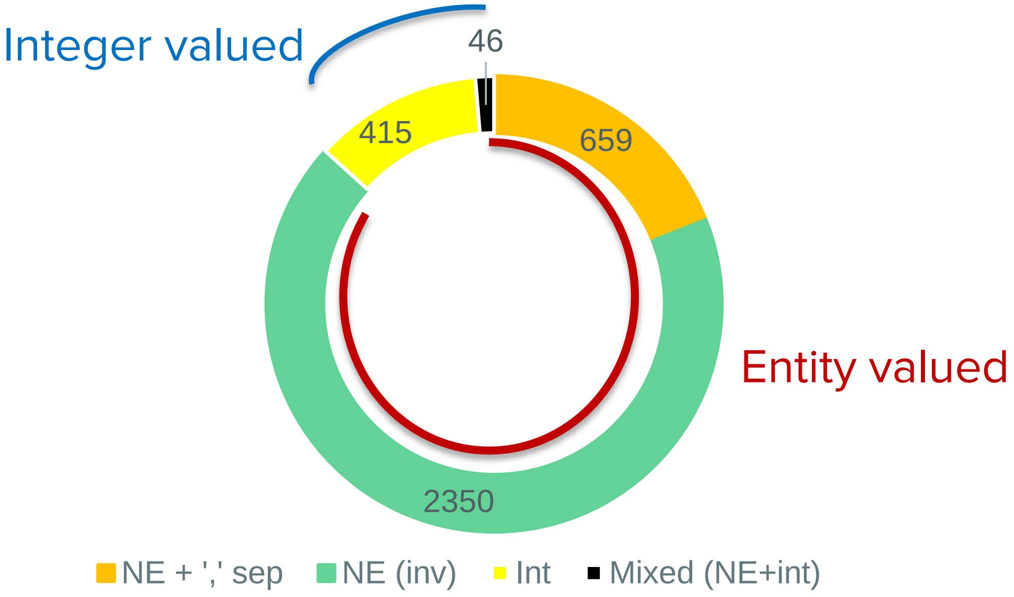
Potential candidates for count related predicates



Integer valued

8. Identification and Extraction - Results

Potential candidates for count related predicates



8. Identification and Extraction - Step 2

Not all candidate predicates are truly count related

Some predicates take both integer values as well as named entities.

Candidate Category	Examples
Entity valued (enumerable?)	nationality, child, debutTeam, gender
Mixed	children, employees
Integer values (enumerating?)	population, numberOfStudents/Employees, floorCount
Negatives (other)	Elevation, foundingDate

Solution: Classifier on top of candidates

8. Identification and Extraction - Results

- Used 56 annotated predicates out of 659 NE valued clean predicates
- Features used:
  - Interest Ratio of predicate phrase in plural over singular form based on frequency of occurrence in Google search
  - N-gram frequency of predicates in text co-occurring with count information
    - Relative frequency of “2/Two/No. of <predname>” over “<predname>”
  - Average number of object entities that a predicate takes per subject
  - Predicate frequency over the entire database
- Result: 0.73 accuracy, 0.7 precision, 0.97 recall
  - To be taken with caution, data small

8. Identification and Extraction - Results

- Used 50 annotated predicates out of 415 integer valued clean predicates
- Features used:
  - Subject Type of predicate phrase may belong to Organisation, Work, Person
  - Maximum object value of the predicates per subject
- Result: 0.86 accuracy, 0.78 precision, 0.91 recall
  - To be taken with caution, data small

Mixed and inverse predicates not classified.

9. Predicate Alignment

Approach:

1. Frequency of co-occurrence
    a. Absolute, Jaccard, relative overlap, pairwise mutual information

Relatedness [Absolute(PMI) scores]		Entity valued predicates (NE+’,’sep, NE(inv), Mixed)		
		institutions^{-1}	chancellor	artist
Integer valued predicates (Int, Mixed)	facultySize	8.13 (148)	7.35 (224)	0
	musicVideos	0	0	5.87 (2127)

9. Predicate Alignment

Approach:

1. Frequency of co-occurrence
   a. Absolute, Jaccard, relative overlap, pairwise mutual information

Relatedness [Absolute(PMI) scores]		Entity valued predicates (NE+’,’sep, NE(inv), Mixed)		
		institutions^{-1}	chancellor	artist
Integer valued predicates (Int, Mixed)	facultySize	8.13 (148)	7.35 (224)	0
	musicVideos	0	0	5.87 (2127)

9. Predicate Alignment

Approach:

1. Frequency of co-occurrence
   a. Absolute, Jaccard, relative overlap, pairwise mutual information

Relatedness [Absolute(PMI) scores]		Entity valued predicates (NE+’,’sep, NE(inv), Mixed)		
		institutions^{-1}	chancellor	artist
Integer valued predicates (Int, Mixed)	facultySize	8.13 (148)	7.35 (224)	0
	musicVideos	0	0	5.87 (2127)

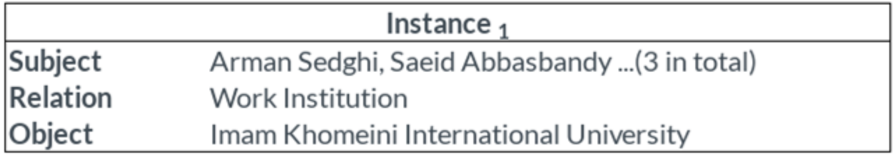
9. Predicate Alignment

2. How frequently #entities supported by an entity valued predicates almost matches the object value of an integer valued predicate for the same subject?
   a. count(children) ≂ numberOfChildren for the co-occurring subjects

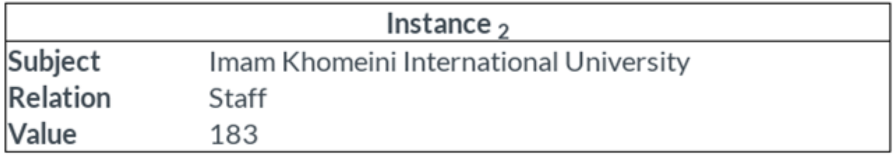
Enumerable	Enumerating	#co-occurring subjects	Mean match	90 percentile #instances	90 percentile value
Bishop	Number of members	13	0.00013	3	415475.4
Institution	Faculty size	63	0.008	8.8	5648
Executive producer	Number of seasons	2149	0.55	5	6

10. Alignment Evaluation

- Crowdsource ground truth for 75 clean, classified and aligned Enumerating and Enumerable predicate pairs
- Top 5 enumerating predicates for 15 enumerable predicates ranked by PMI scores
- Pairs rated on **topical relevance** (same, related and unrelated topics) and **quantification** (exact, inexclusive, related ,unrelated)
- 3 opinions on each pair



Instance $_1$	
Subject	Arman Sedghi, Saeid Abbasbandy ...(3 in total)
Relation	Work Institution
Object	Imam Khomeini International University



Instance $_2$	
Subject	Imam Khomeini International University
Relation	Staff
Value	183

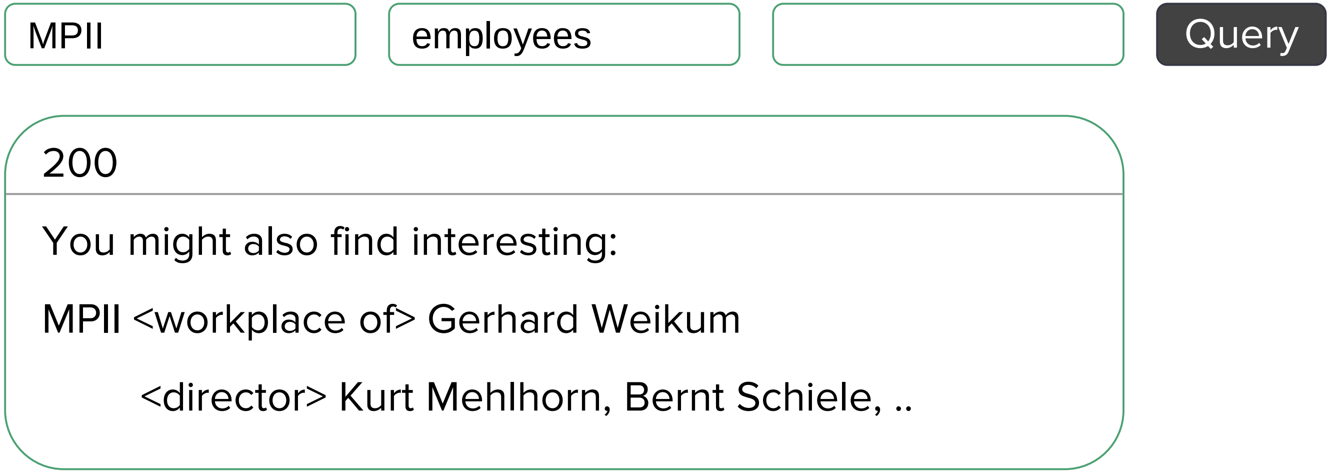
10. Alignment Evaluation

GT Ranking

Enumerable/ Rank	Work Institution	Child organisation	Composer
1	Faculty size, number of undergraduates	Stations	Compilation
2		Fleet, employees	Music videos, eurog
3	Number of postgraduates, staff		
4		Number of employees, routes	Live, certmonth
5	Number of students		

11. Extensions (within scope of thesis)

1. Generalisability - Ongoing work on Wikidata predicates
   a. No problem of mixed predicates
   b. Smaller predicate space
2. Applicability - a simple interface to demonstrate top aligned predicates



MPII

employees

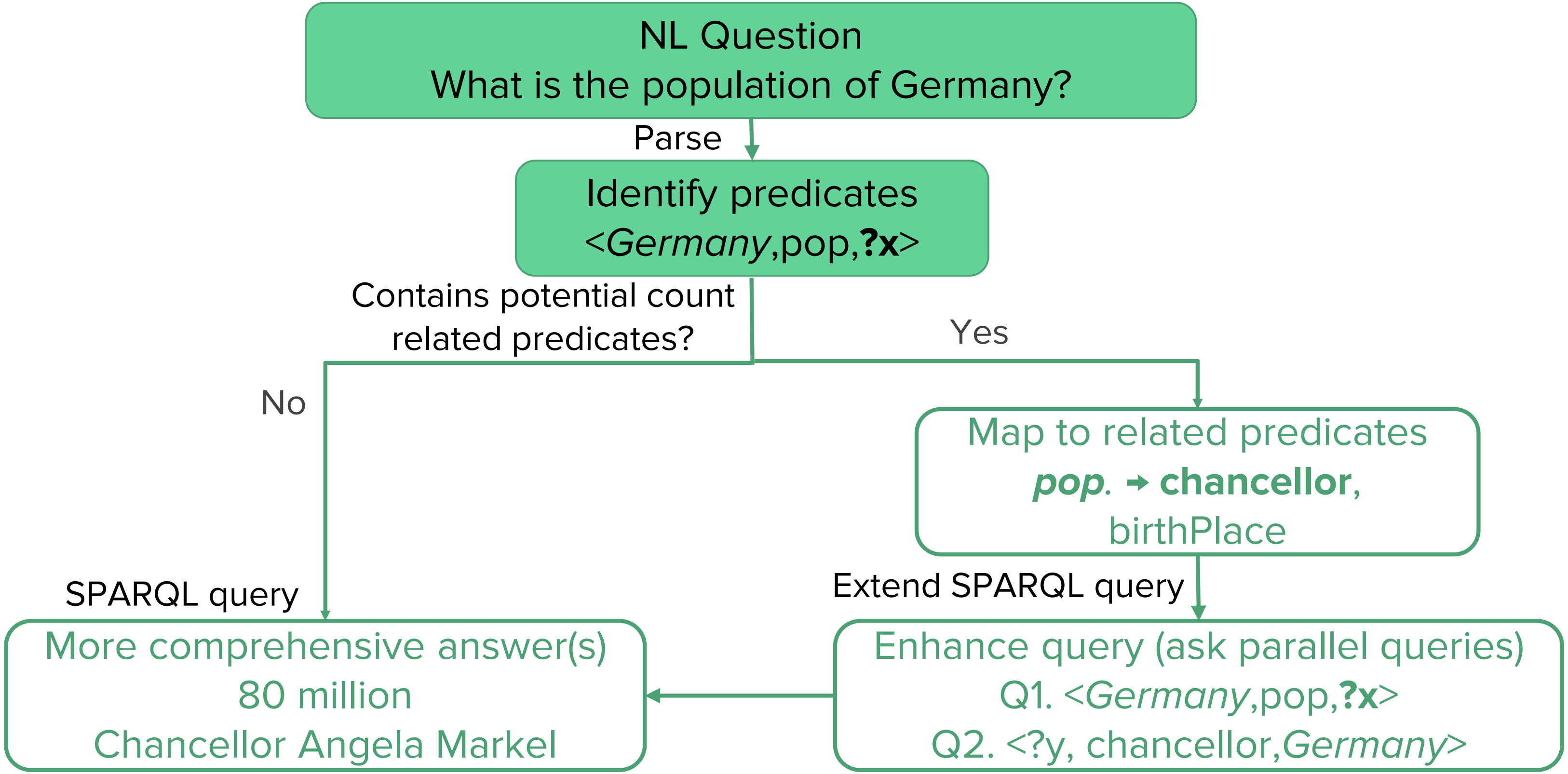
200

You might also find interesting:

MPII <workplace of> Gerhard Weikum

<director> Kurt Mehlhorn, Bernt Schiele, ..

11. Extension 3- Upgrading QA Systems



NL Question

What is the population of Germany?

Parse

Identify predicates

<*Germany*,pop,**?x**>

Contains potential count related predicates?

Yes

Map to related predicates

pop. → **chancellor**, birthPlace

Extend SPARQL query

Enhance query (ask parallel queries)

Q1. <*Germany*,pop,**?x**>

Q2. <?y, chancellor,*Germany*>

No

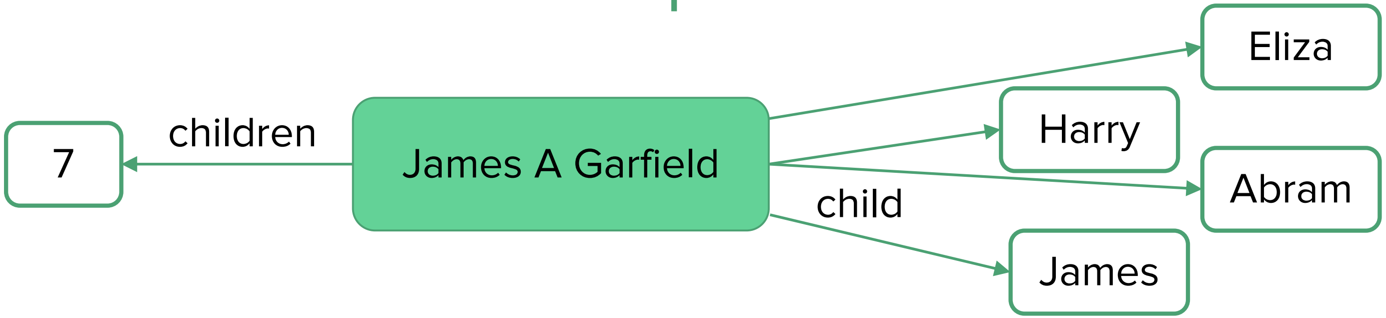
SPARQL query

More comprehensive answer(s)

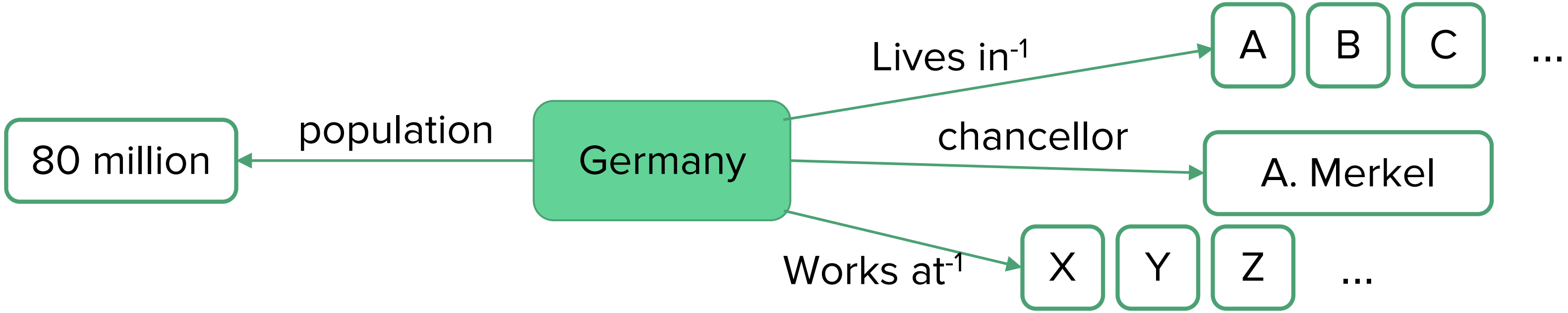
80 million

Chancellor Angela Markel

11. Extension 4: KB Completeness



3 missing children in KB?



Population upper bounded by count value?

12. Summary

1. KBs are a mix of count and fact predicates
2. Predicates take more than one datatypes and hence require fuzzy categorization, filtering and classification
3. Enumerable predicates composed of NE + ’,’ separated, NE(inv)and mixed
4. Enumerating predicates comprise Int and mixed
5. Alignment based on co-occurrence and count values
6. Evaluate by comparing score and GT ranks
7. Extensions covering generalisability and applicability

Thank You!

1. KBs are a mix of count and fact predicates
2. Predicates take more than one datatypes and hence require fuzzy categorization, filtering and classification
3. Enumerable predicates composed of NE + ’,’ separated, NE(inv)and mixed
4. Enumerating predicates comprise Int and mixed
5. Alignment based on co-occurrence and count values
6. Evaluate by comparing score and GT ranks
7. Extensions covering generalisability and applicability