Count Information in KBs and Text

Shrestha Ghosh Simon Razniewski Gerhard Weikum

Talk at DIG group 29.07.2021

About Me

PhD (2019-)

- Databases and Information Systems group in MPI, Informatics
- Saarland Informatics Campus, Saarbrücken

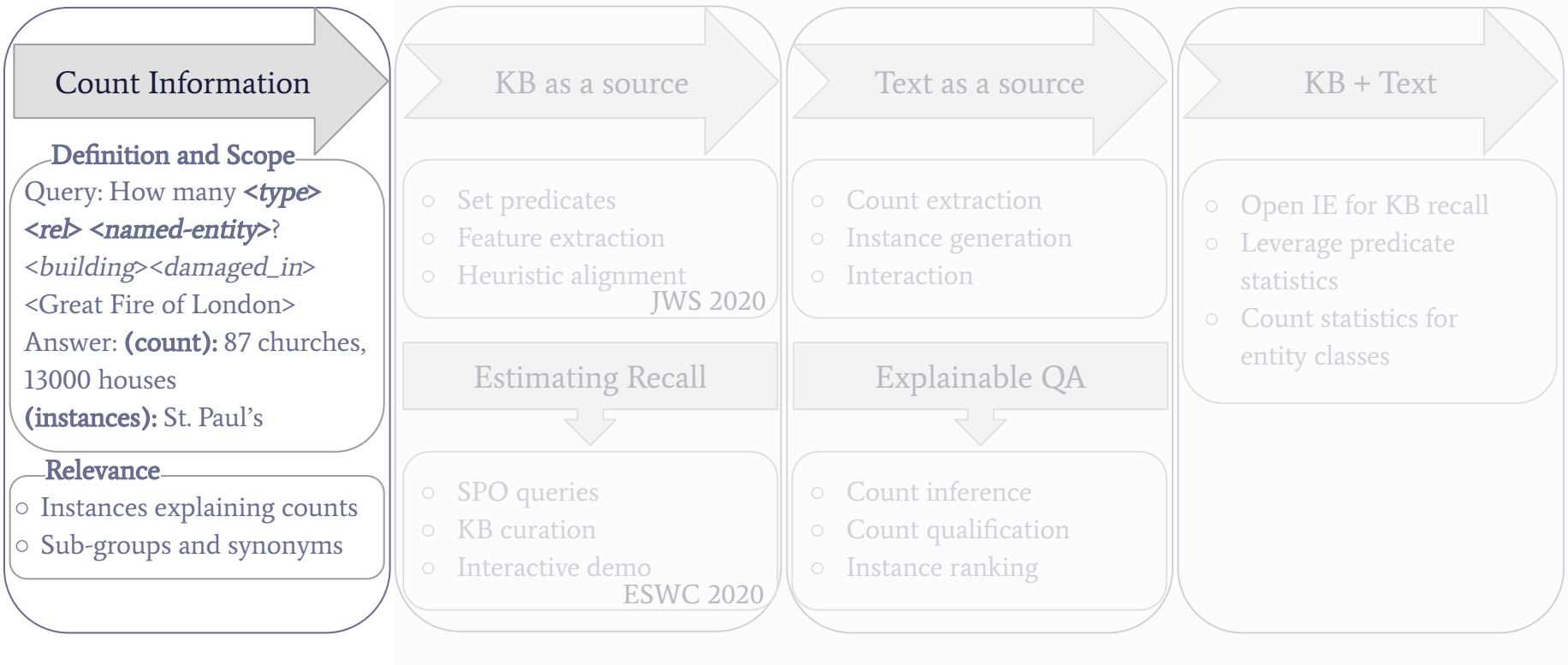
MSc (2019)

- Saarland Informatics Campus, Saarbrücken

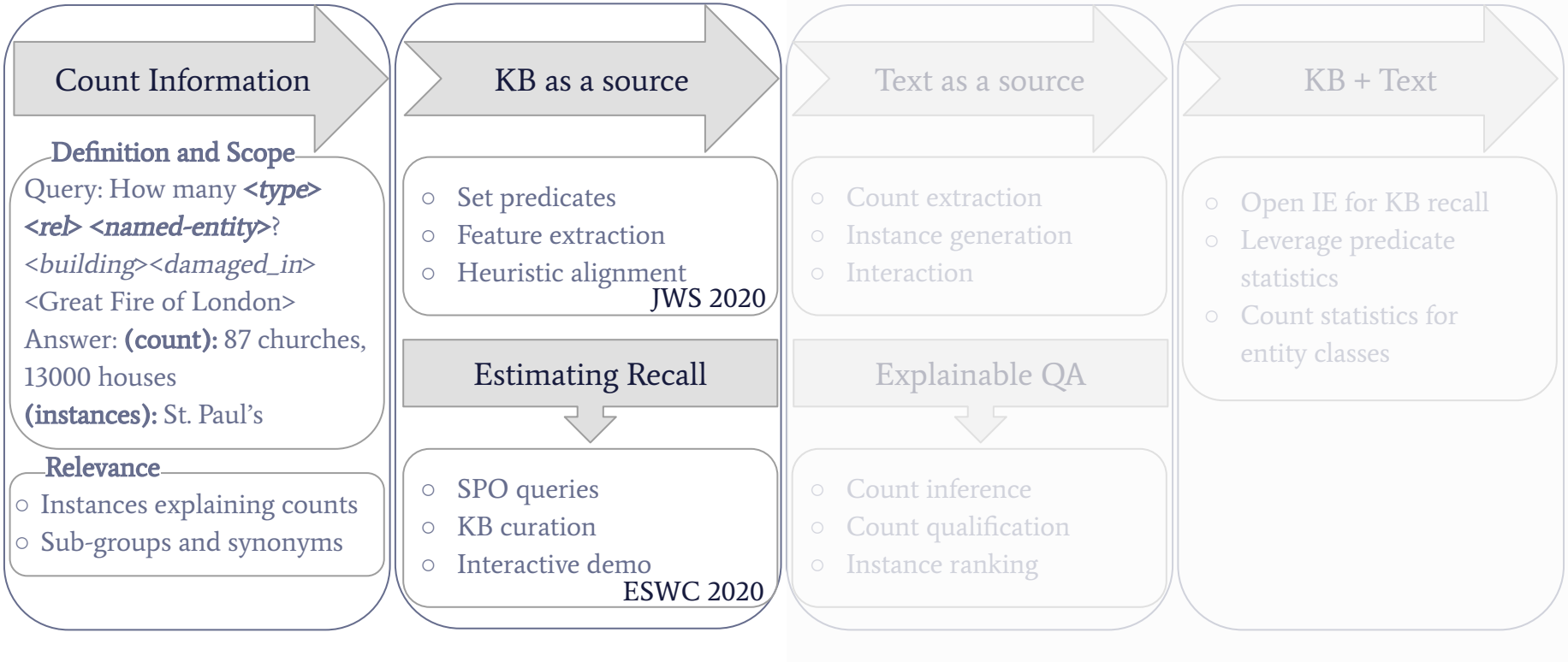
BTech (2016)

- IIEST, Shibpur (India)

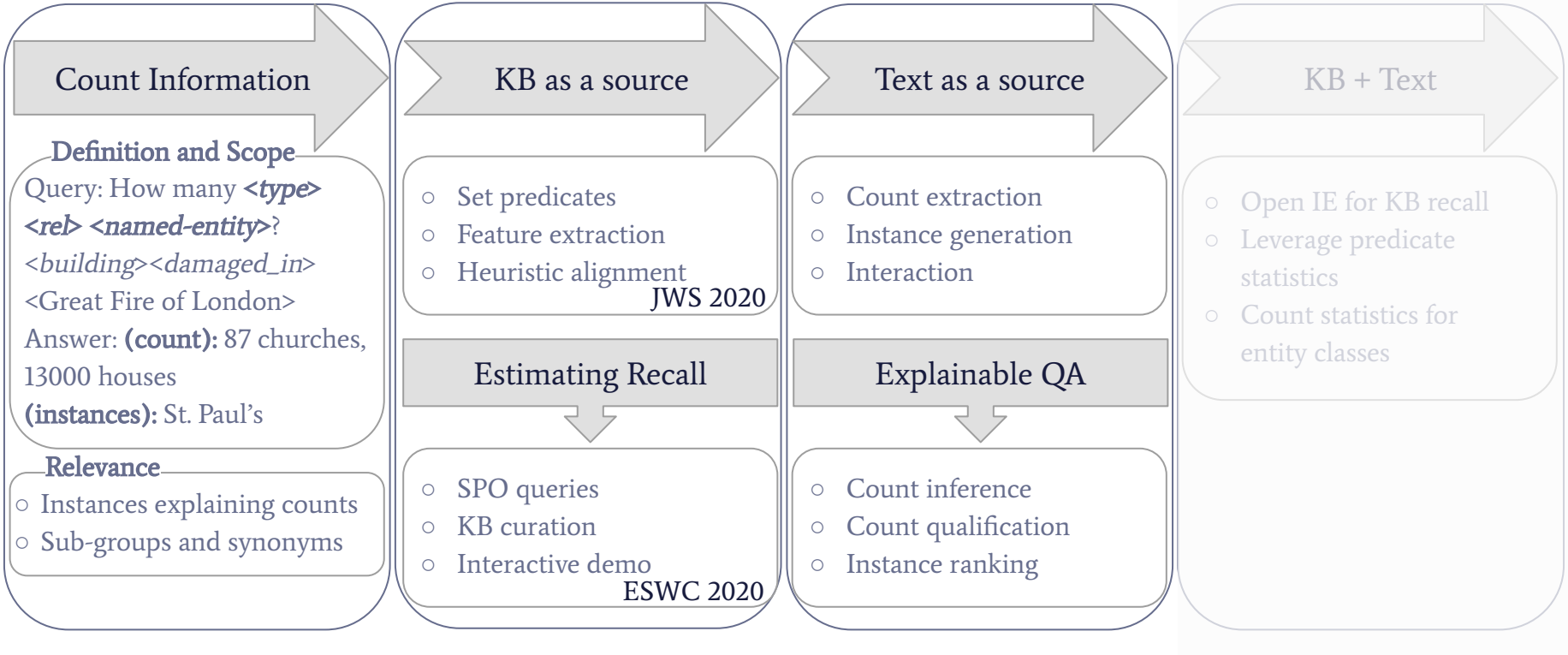
Research Summary



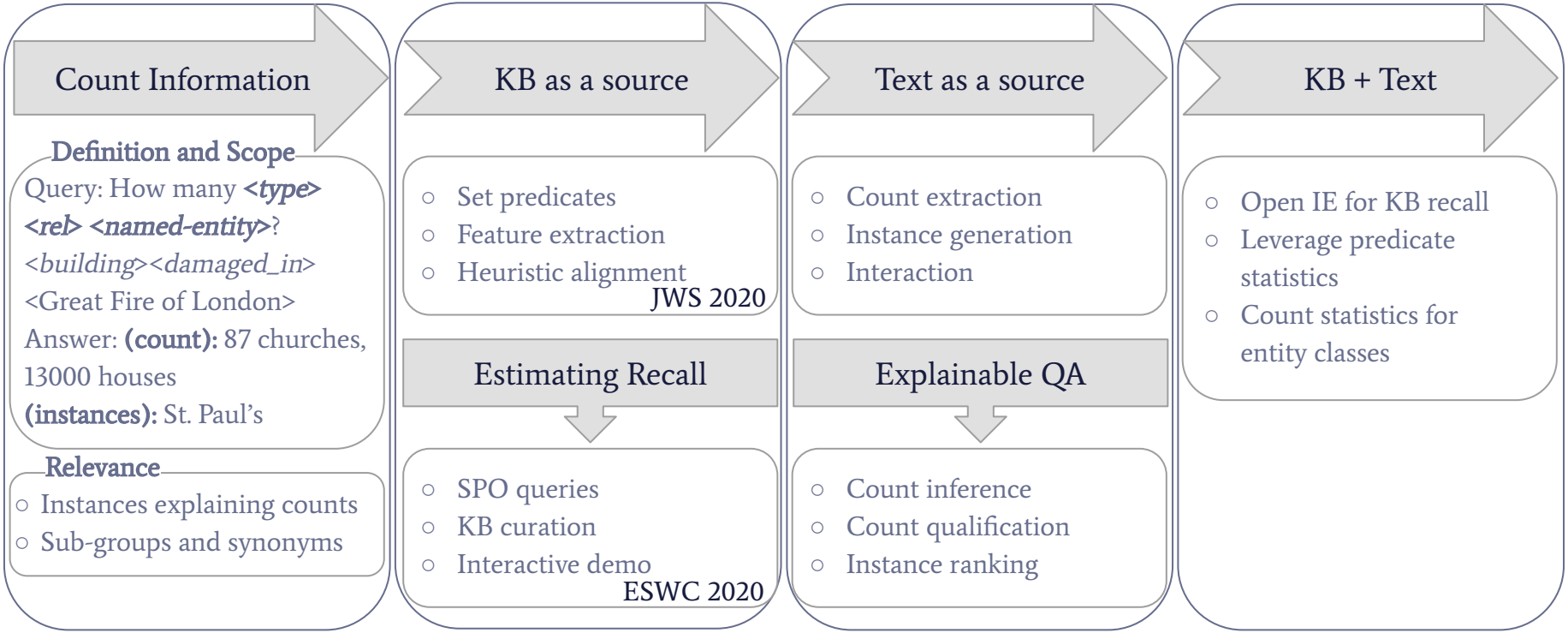
Research Summary



Research Summary



Research Summary



What is count information?

Relation between an entity and a set of entities

Saarland
University

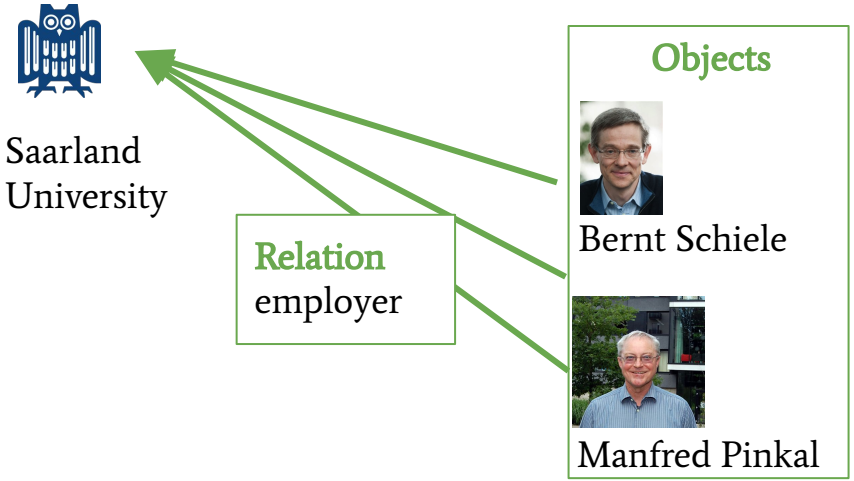
Objects

Bernt Schiele

Manfred Pinkal

What is count information?

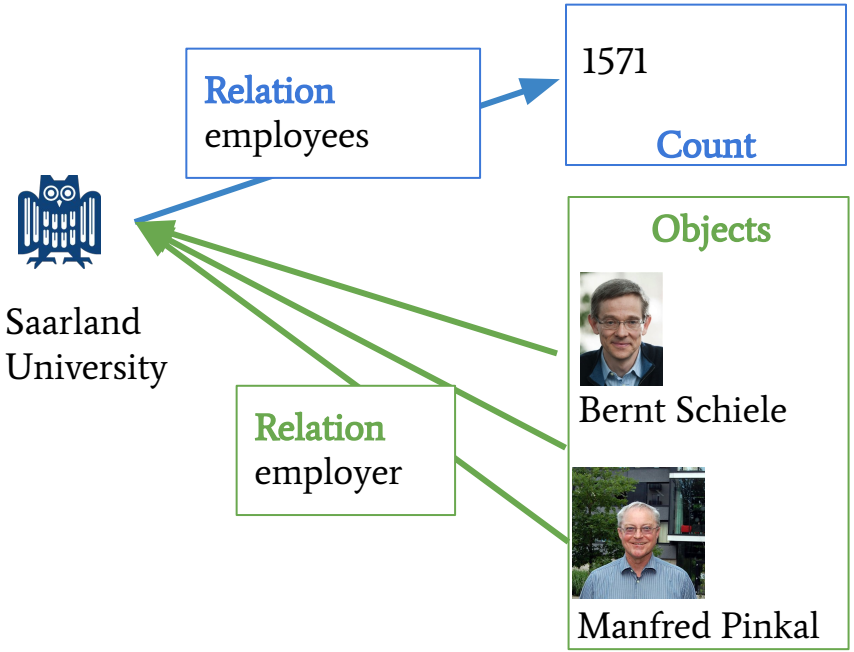
Relation between an entity and a set of entities



Expressed as **entities** or objects in the set

What is count information?

Relation between an entity and a set of entities



Expressed as **count** or cardinality of the set

Expressed as **entities** or objects in the set

Why do we need count information?

Only counts

(Saarland_University, employees, ?y)

Gives no insight about the entities

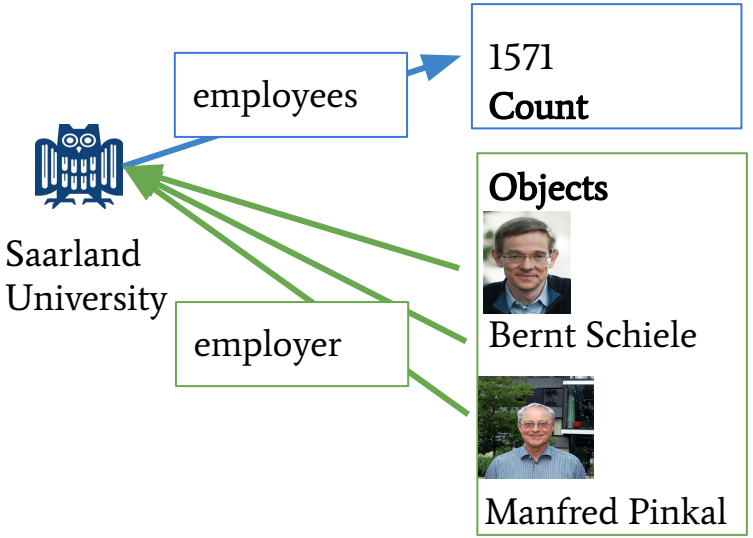
Only entities

(?x, employer, Saarland_University)

May return only a handful of names

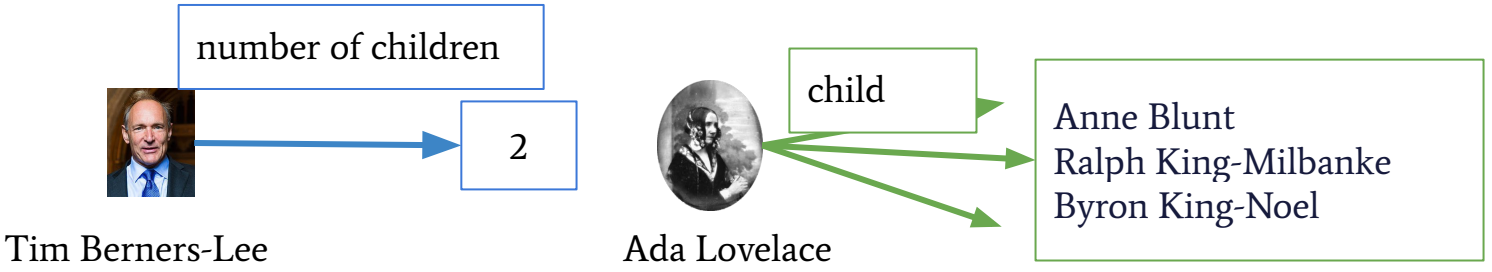
Incomplete positives can benefit from complete counts

Counts can benefit from representative entities



Why do we need count information?

KB mixes counts with standard facts



Why do we need count information?

Analysing KB recall

Noam Chomsky

child → Aviva Chomsky

number of children → 3

Tim Berners-Lee

child → -

number of children → 2

Lines of approach

- IE of count information for KB curation or recall assessment
- Analysing count information in KB
- Analysing count information for Question Answering task

Lines of approach

- IE of count information for KB curation or recall assessment
- Analysing count information in KB
- Analysing count information for Question Answering task

Mirza et al. “Enriching Knowledge Bases with Counting Quantifiers” ISWC 2018

Mirza et al. “Cardinal Virtues: Extracting Relation Cardinalities from Text” ACL 2017

Lines of approach

- IE of count information for KB curation or recall assessment
- Analysing count information in KB
- Analysing count information for Question Answering task

Ghosh et al. “Uncovering hidden semantics of set information in knowledge bases” JWS 2020

Ghosh et al. “CounQER: A System for Discovering and Linking Count Information in Knowledge Bases” ESWC 2020

Lines of approach

- IE of count information for KB curation or recall assessment
- Analysing count information in KB
- Analysing count information for Question Answering task

- Low entry barrier for creating an automated training dataset
- No need to rely on incomplete KB for gold standard counts
- Consolidate results from multiple text sources

What's in a Count?

Problem: Answering count queries with explanations

Input:

- A count query q
- A set of relevant documents D

Task: Determine a consolidated count to the query q with notable instances that instantiate the count

What's in a Count?

Count Query

How many buildings were damaged in the Great Fire of London?

<type> *<relation>* *<named-entity>*

What's in a Count?

Count Query

How many buildings were damaged in the Great Fire of London?
<type> *<relation>* *<named-entity>*

Tasks

- **Consolidation** over count distribution from multiple sources.
- **Count qualification** for subgroups and synonyms.
- **Notable instances** with evidence which instantiate the counts.

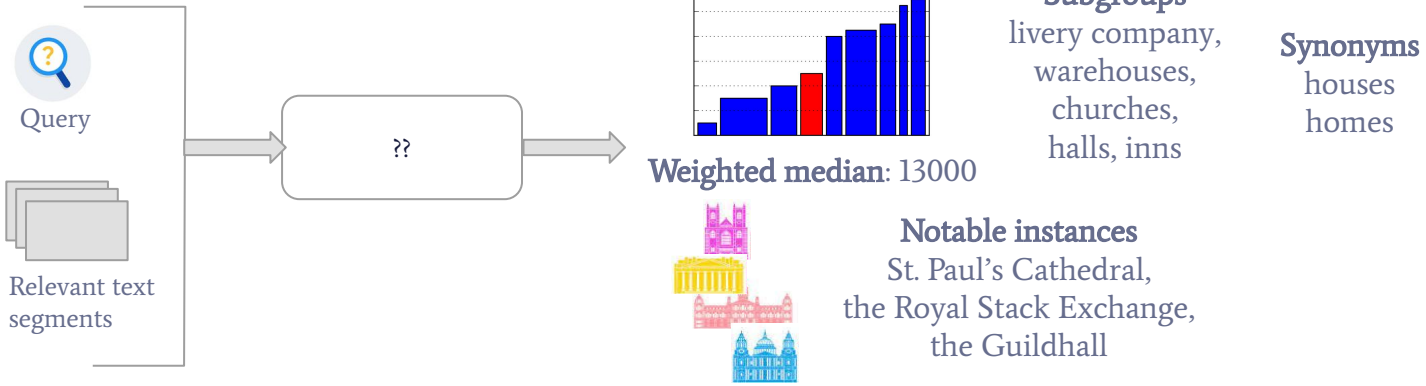
What's in a Count?

Count Query

How many buildings were damaged in the Great Fire of London?
<type> *<relation>* *<named-entity>*

Tasks

- **Consolidation** over count distribution from multiple sources.
- **Count qualification** for subgroups and synonyms.
- **Notable instances** with evidence which instantiate the counts.



Existing Paradigms and their limitations

KB QA

- Low KB recall
- Aggregations on list (QAnswer, Diefenbach et al. ESWC 2020)

Textual QA

- Extractive QA (Dense Passage Retrieval, Karpukhin et al. EMNLP 2020; DROP, Dua et al. NAACL 2019)
- Ranked answer spans without any consolidation

KB+Textual QA

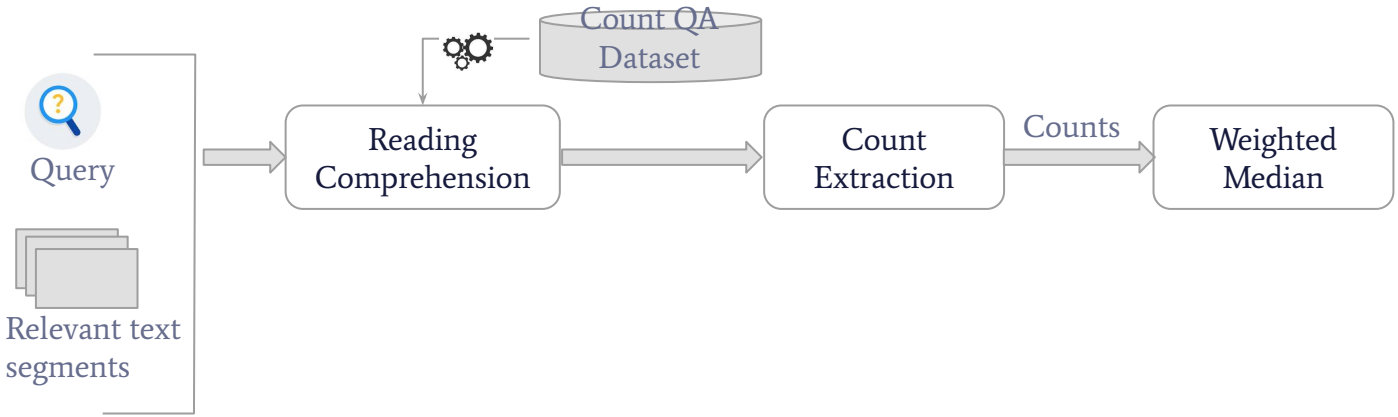
- Search Engines have high precision and low recall esp. on tail entities
- Hybrid QA systems

What's Missing?

Answer consolidation

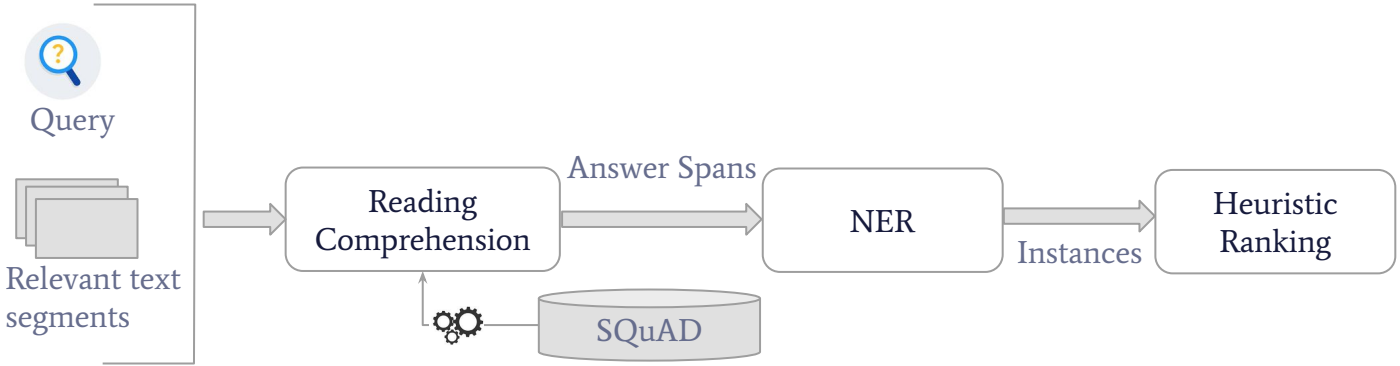
1. Expand answer scope to allow multiple correct answers, estimates or ranges for counts.
2. Explainability
   a. instances can be useful to explain counts
   b. counts themselves are multifaceted - synonyms, sub-groups

Approach



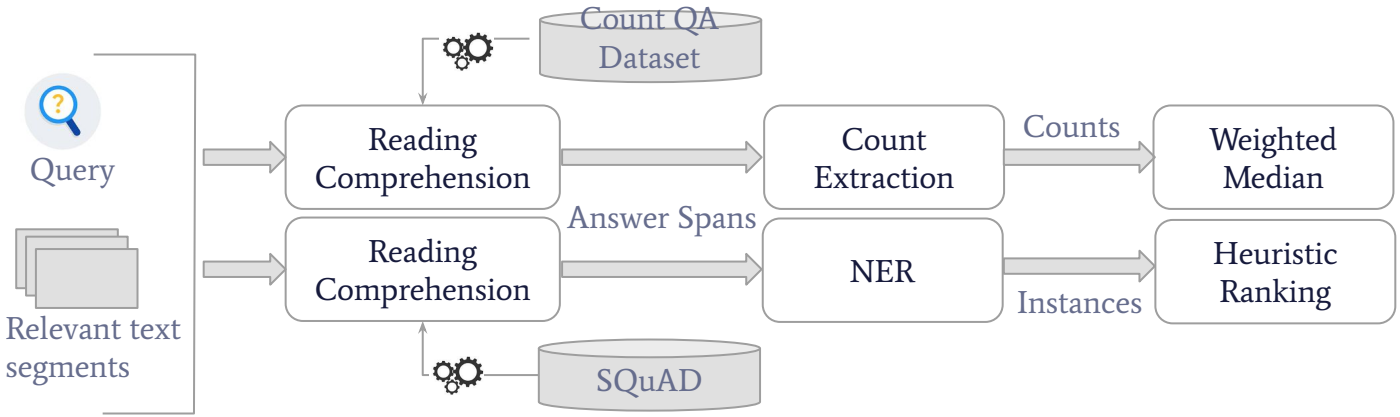
Count Inference

Approach



Notable Instances

Approach



NLCounQER

Results

	Generating Instances				Count Inference	
	MAP@10	AR@10	MAP@20	AR@20	P	RP
Frequent	0.053	0.017	0.055	0.026	-	-
Type-compatible	**0.206**	0.077	**0.177**	0.126	-	-
DistilBERT [34]	-	-	-	-	0	0
SpanBERT [18]	-	-	-	-	**0.35**	0.4
NLCounQER (Span-predicted)	0.157	**0.198**	0.153	**0.231**	**0.35**	**0.45**

Comparison of NLCounQER with different count inference and instance generation baselines on Stresstest queries and search engine snippets

Demo system: nlcounqer.mpi-inf.mpg.de

Challenges

- Lack of annotated data
  - Training for count contexts
  - Evaluation data
- Extractive QA is a black box
  - Is it really learning to predict count spans?
  - Multi-span prediction is underexplored
- Getting entities from text without linking is difficult
  - Can Wikipedia hyperlinks help?

Possible directions and challenges

- Count information extraction from text
  - evidence of count of islands in hawaii (~140) >> KB entities (22)
  - use count queries and count predicates as starting points for harvesting

Possible directions and challenges

- Count information extraction from text
  - evidence of count of islands in hawaii (~140) >> KB entities (22)
  - use count queries and count predicates as starting points for harvesting
- Estimating KB count recall
  - entity level - alignment inconsistencies
  - class level - all humans who have *number_of_children* should have *child* and vice-versa

Possible directions and challenges

- Count information extraction from text
  - evidence of count of islands in hawaii (~140) >> KB entities (22)
  - use count queries and count predicates as starting points for harvesting
- Estimating KB count recall
  - entity level - alignment inconsistencies
  - class level - all humans who have *number_of_children* should have *child* and vice-versa (children in Wikidata, inconsistency vs. incompleteness)
- Extending KB with count information
  - entities not yet present in KB
  - contentious counts - collection size of a museum vs visitors per year (examples: entity, query)

Questions?

- Count information is the relation between an entity and a set of entities
  - represented as a count, or
  - as an individual entity from the set
- Count information can help in KB curation and QA tasks
- Count queries can be enhanced using consolidated counts and entities