Count Information: Retrieving and Estimating Cardinality of Entity Sets from the Web

Shrestha Ghosh

Doctoral Thesis Defense

Outline

1. Introduction
2. Contributions
    - RQ1: Identifying and Aligning Count Information in KBs
    - RQ2: Aggregating Count Information from Web Snippets
    - RQ3: Cardinality Comparison
    - RQ4: Cardinality Estimation
3. Summary
4. Limitations & Future Work
5. Conclusion

1. Introduction

Introduction: Background

Count Information captures the **cardinality** of a **set of entities**, directly as a **count** or as **enumerations**.

	Example 1	**Example 2**
Set of Entities	Nobel Prize Winners in Physics	Physicists in the world
Cardinality / Count	224	~1 million
Enumeration	{Max Planck, Marie Curie, … , }	{Jocelyn Burnell, Peter Higgs, …, }

Ground-truth cardinality of Nobel Prize Winners obtained from NobelPrize(2023) and physicists from Day(2015). Enumerations are obtained from Wikidata, last accessed 27.09.2024.

Introduction: Background

Count Information captures the **cardinality** of a **set of entities**, directly as a **count** or as **enumerations**.

	Example 1	**Example 2**
Set of Entities	Nobel Prize Winners in Physics *crisp, popular*	Physicists in the world *fuzzy, large*
Cardinality / Count	224 *exact*	~1 million *estimate*
Enumeration	{Max Planck, Marie Curie, … , } *complete (224 entities)*	{Jocelyn Burnell, Peter Higgs, …, } *incomplete (33,000 entities)*

Ground-truth cardinality of Nobel Prize Winners obtained from NobelPrize(2023) and physicists from Day(2015). Enumerations are obtained from Wikidata, last accessed 27.09.2024.

Introduction: Knowledge on the Web

1. **Explicit Knowledge**

   Entity-centric KB

Suchaneck et al. (2007), Auer et al. (2007), Vrandečić (2012)

Singhal (2012), Qian (2013), Baidu (2020)

Introduction: Knowledge on the Web

1. **Explicit Knowledge**

   Entity-centric KB

   Aspect-based

Numeric

Saha et al. (2017), Ho et al. (2022)

Negative

Darari et al. (2015), Arnaout et al. (2021)

Temporal

Hoffart et al.(2013), Jain et al. (2020)

Commonsense

Liu and Singh (2004), Tandon et al. (2014), Sap et al. (2019), Romero et al. (2019), Nguyen et al. (2021)

Introduction: Knowledge on the Web

1. **Explicit Knowledge**

   Entity-centric KB

   Aspect-based

2. **Parametric Latent Knowledge**

   Transformer-based language models

GPT4-o and DALL-E AI

LLMs

Radford et al. (2018), Devlin et al. (2019), Raffel et al. (2020), Du et al. (2022)

KBs from LLM

Petroni et al. (2019), Singhania et al. (2023)

Commonsense

Elazar et al. (2019)

Retrieval-augmented generation

Lewis et al. (2020), Izacard and Grave (2021)

Introduction: Accessing Count Information

how many Nobel Prizes won by Marie Curie

Marie Curie / Nobel Prize / Wins / **Count**

2

from KB

Introduction: Accessing Count Information

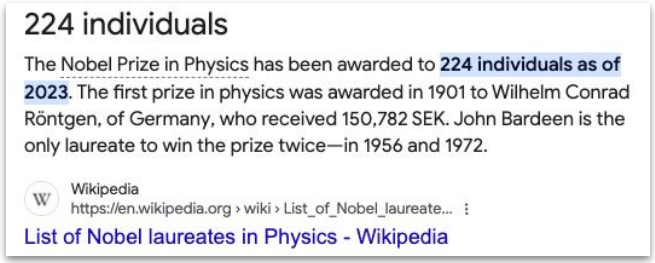
how many Nobel Prizes won by Marie Curie

Marie Curie / Nobel Prize / Wins / **Count**

2

from KB

how many Nobel Prize winners in Physics



from featured snippet

Introduction: Accessing Count Information

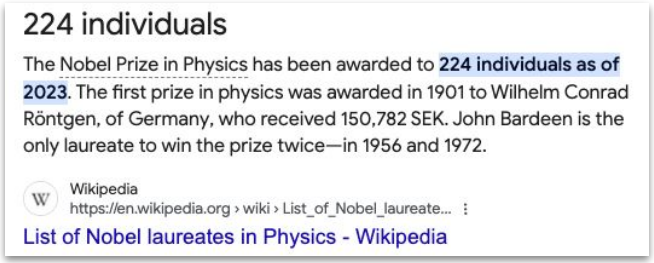
how many Nobel Prizes won by Marie Curie

Marie Curie / Nobel Prize / Wins / **Count**

2

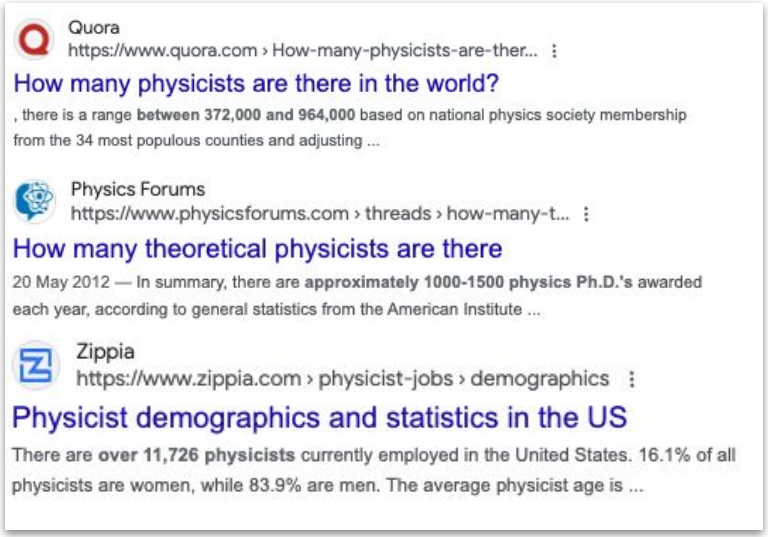
from KB

how many Nobel Prize winners in Physics



from featured snippet

how many physicists are there in the world



no direct answers

Introduction: Accessing Count Information

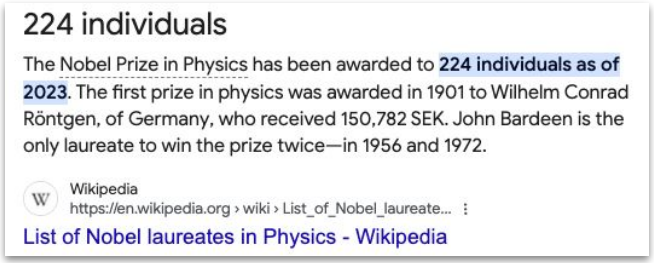
how many Nobel Prizes won by Marie Curie

Marie Curie / Nobel Prize / Wins / **Count**

2

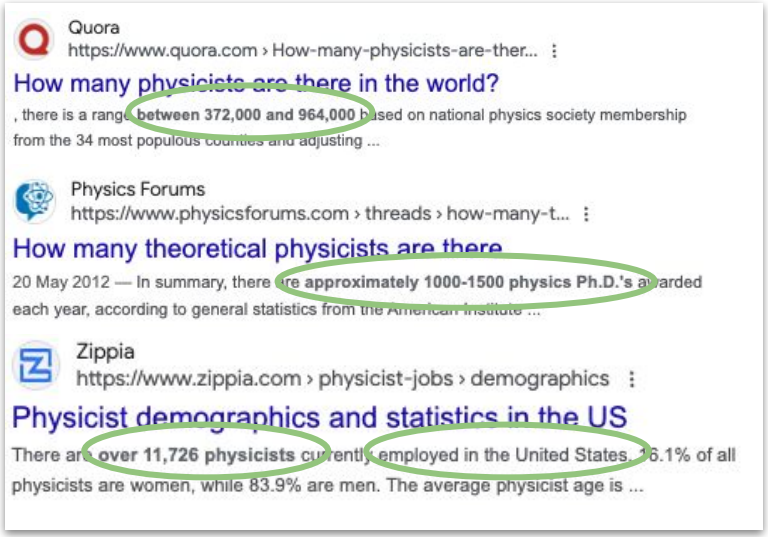
from KB

how many Nobel Prize winners in Physics



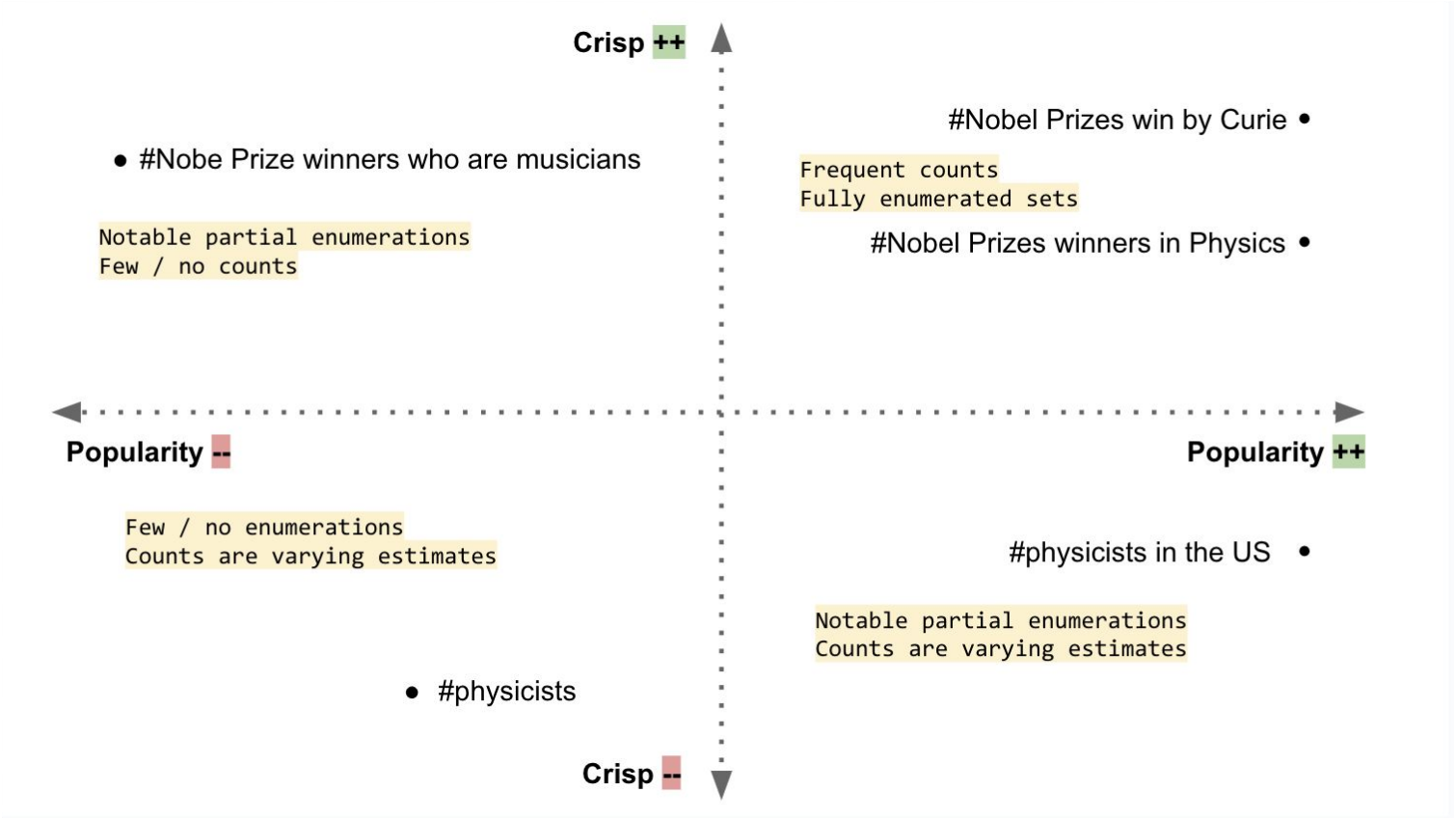
from featured snippet

how many physicists are there in the world

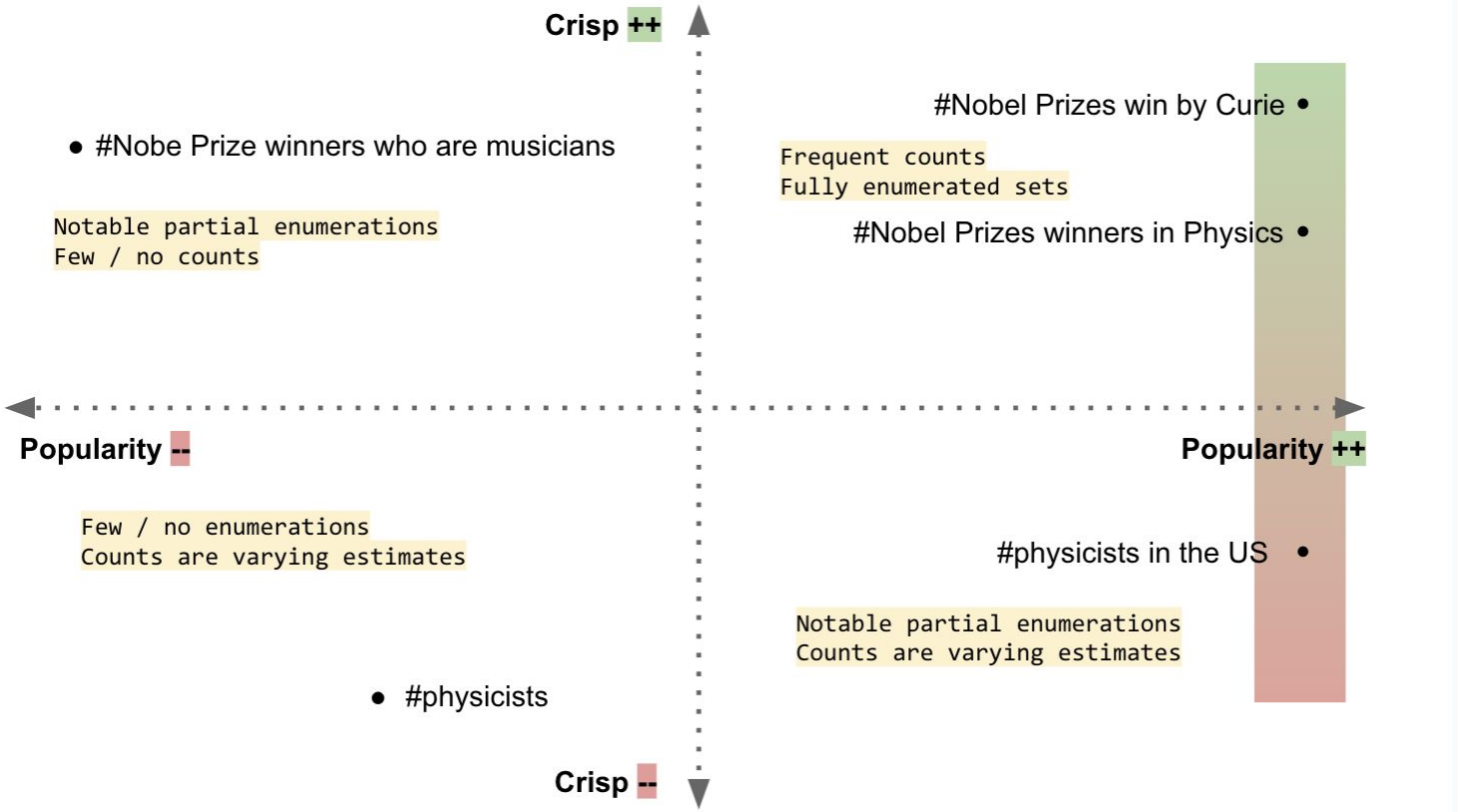


no direct answers

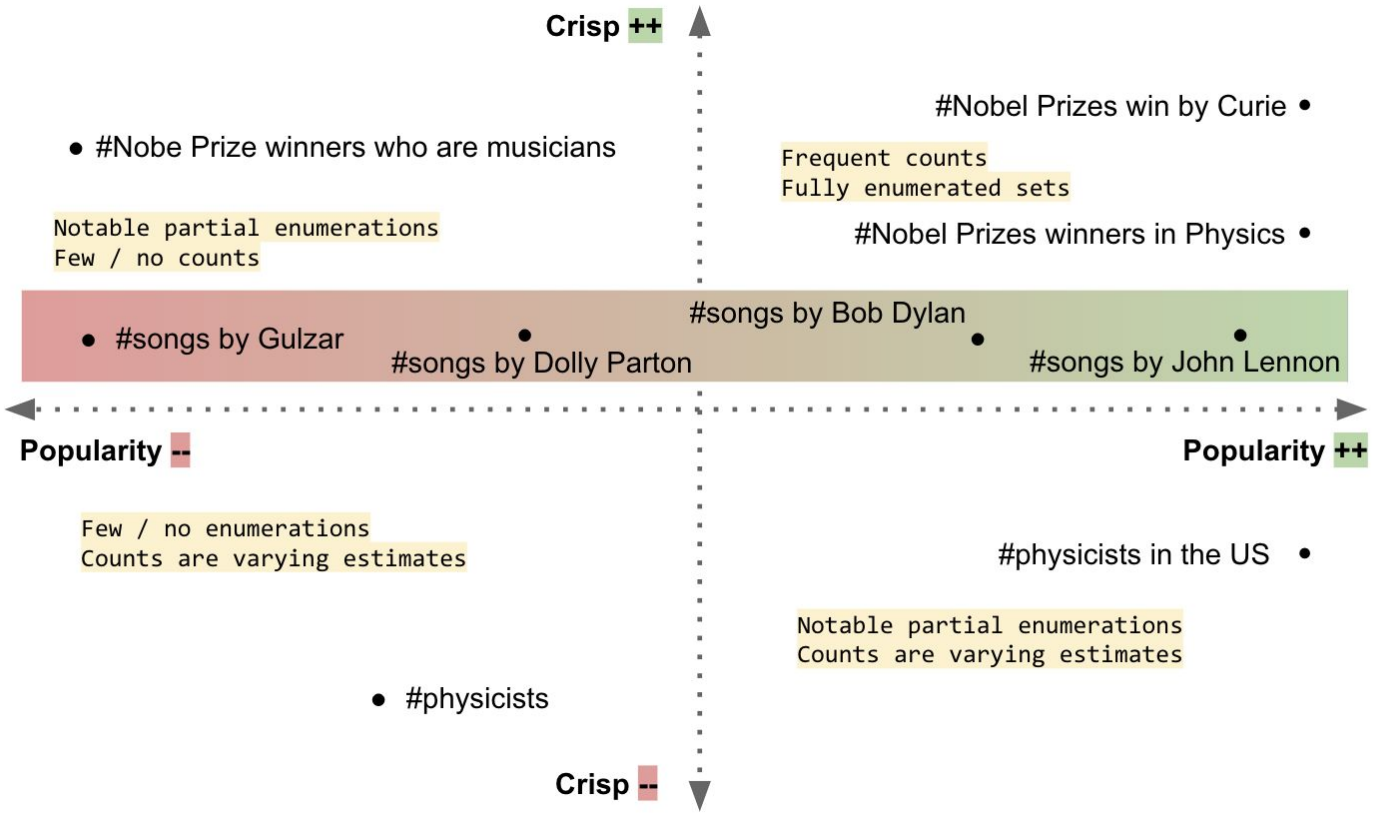
Introduction: Coverage Bias



Introduction: Coverage Bias



Introduction: Coverage Bias



Introduction: Count Question Answering

5%-10% TREC QA datasets contain counts

Mirza et al. (2018)

Count questions not dealt in a principled manner.

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Low recall

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Low recall

Open QA focussed on single best answer

Karpukhin et al. (2020), Izacard and Grave (2021), Yu et al. (2022)

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Low recall

Open QA focussed on single best answer

Karpukhin et al. (2020), Izacard and Grave (2021), Yu et al. (2022)

Ignores underlying distribution

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Low recall

Open QA focussed on single best answer

Karpukhin et al. (2020), Izacard and Grave (2021), Yu et al. (2022)

Ignores underlying distribution

Benchmarks: exact match, n-gram overlap

SQuAD [Rajpurkar et al. 2016, 2018], LC-QuAD 2.0 [Dubey et al. (2019)]

Introduction: Count Question Answering

Count questions not dealt in a principled manner.

KB QA ad-hoc aggregate over objects

Bast and Hausmann (2015), Diefenbach et al. (2018)

Low recall

Open QA focussed on single best answer

Karpukhin et al. (2020), Izacard and Grave (2021), Yu et al. (2022)

Ignores underlying distribution

Benchmarks: exact match, n-gram overlap

SQuAD [Rajpurkar et al. 2016, 2018], LC-QuAD 2.0 [Dubey et al. (2019)]

Does not transfer to counts

Introduction: Cardinality Estimation

Cardinality estimation from KBs

Trushkowsky et al. (2013), Galárraga et al. (2017), Soulet et al. (2018), Luggen et al. (2019)

Introduction: Cardinality Estimation

Cardinality estimation from KBs

Trushkowsky et al. (2013), Galárraga et al. (2017), Soulet et al. (2018), Luggen et al. (2019)

Prone to underestimations

Cardinality cues in KB unused

Introduction: Cardinality Estimation

Cardinality estimation from KBs

Trushkowsky et al. (2013), Galárraga et al. (2017), Soulet et al. (2018), Luggen et al. (2019)

Prone to underestimations

Cardinality cues in KB unused

Cardinality estimation from open IE

Mirza et al. (2017), Razniewski et al. (2019)

Introduction: Cardinality Estimation

Cardinality estimation from KBs

Trushkowsky et al. (2013), Galárraga et al. (2017), Soulet et al. (2018), Luggen et al. (2019)

Prone to underestimations

Cardinality cues in KB unused

Cardinality estimation from open IE

Mirza et al. (2017), Razniewski et al. (2019)

Depends on single source

2. Contributions

Contributions: Overview

RQ1

Counts in KBs (CounQER)

*How to **identify** and **align predicates** that store count information?*

JWS 2020,
ESWC 2020 (demo)

Contributions: Overview

RQ1. How can we identify and align count information in KBs?

Aggregating objects of (enumerating) predicates often mislead.

number of children of Charlie Chaplin (GT = 11)

child	Charles Chaplin	Victoria Chaplin
	Geraldine Chaplin	Eugene Chaplin
	Michael Chaplin	Jane Chaplin
	Josephine Chaplin	Christopher Chaplin
(Wikidata)	Sydney Chaplin	

number of employees at Saarland University (GT = 1500)

```
select  count(?s) where {
  ?s
    dbp:workInstitution | dbp:workplaces
  dbr:Saarland_University.
}
```

=18

(DBpedia)

Contributions: Overview

RQ1. How can we identify and align count information in KBs?

number of children	11

We can fall back on cardinality storing (counting) predicates.

dbp:academicStaff	1571

number of children of Charlie Chaplin (GT = 11)

child		
	Charles Chaplin	Victoria Chaplin
	Geraldine Chaplin	Eugene Chaplin
	Michael Chaplin	Jane Chaplin
	Josephine Chaplin	Christopher Chaplin
(Wikidata)	Sydney Chaplin	

number of employees at Saarland University (GT = 1500)

```
select  count(?s) where {
  ?s
    dbp:workInstitution | dbp:workplaces
  dbr:Saarland_University.
}
```

=18

(DBpedia)

Contributions: Overview

RQ1

Counts in KBs (CounQER)

*How to **identify** and **align predicates** that store count information?*

JWS 2020, ESWC 2020 (demo)

RQ2

Counts in Web Snippets (CoQEx)

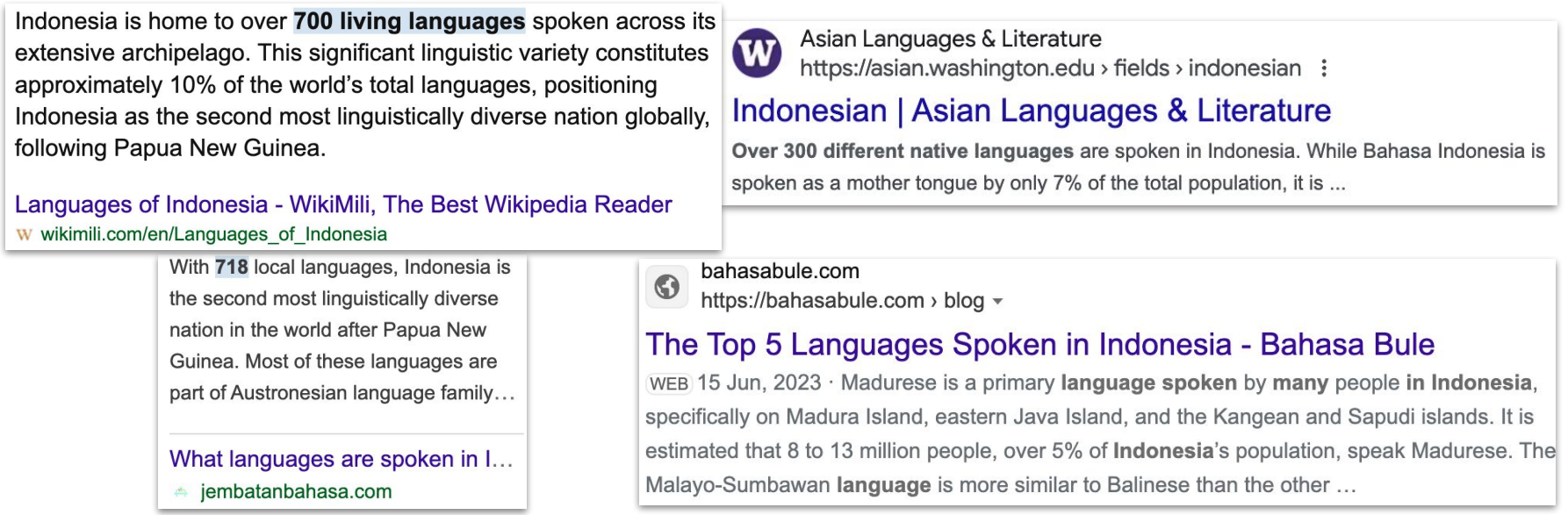
*How to **aggregate counts** and **enumerations** from Web snippets?*

SIGIR 2022, JoWS 2022 WSDM 2023 (demo)

Contributions: Overview

RQ2. How can we aggregate count information from Web snippets?

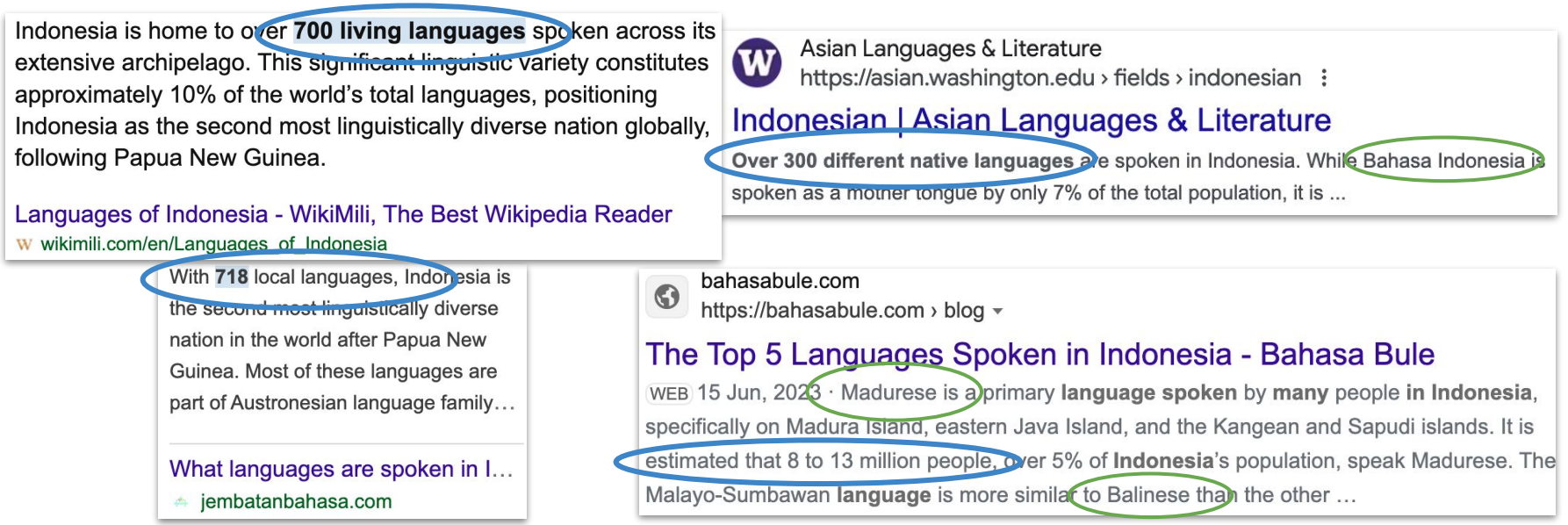
How many languages are spoken in Indonesia?



Contributions: Overview

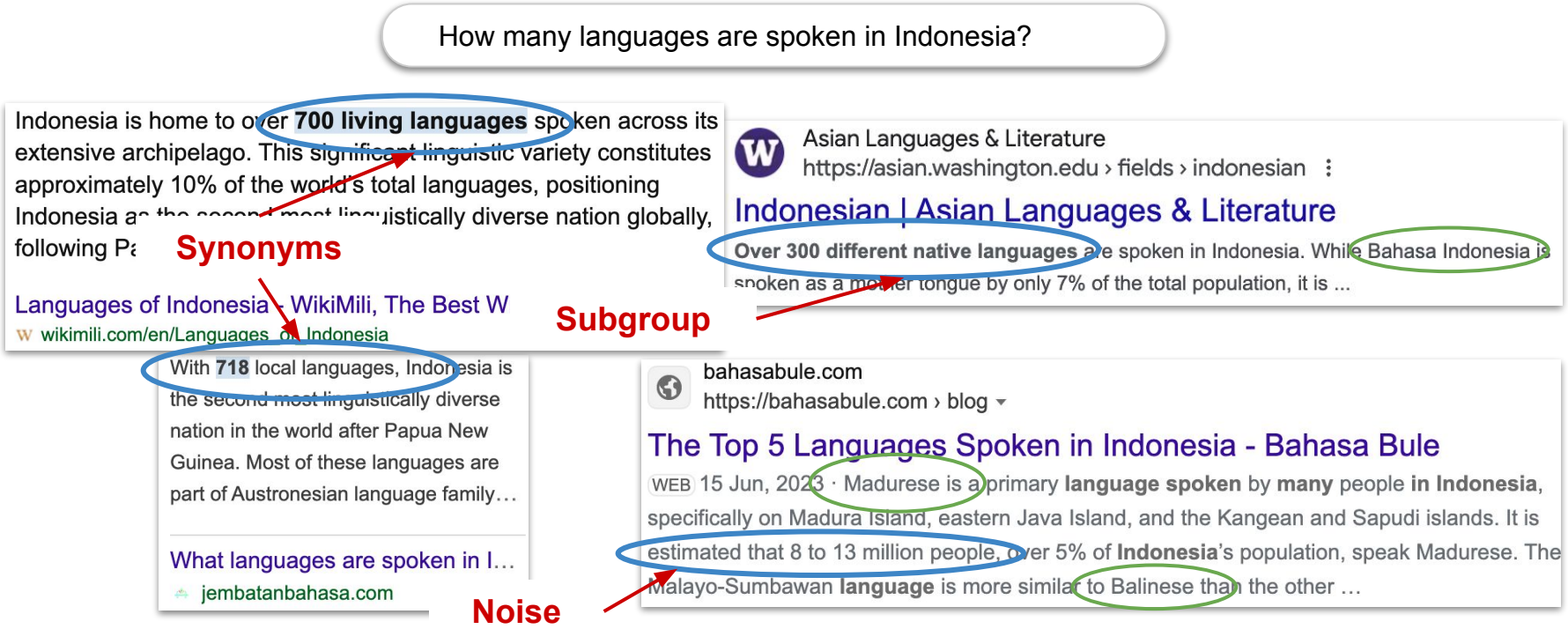
RQ2. How can we aggregate count information from Web snippets?

How many languages are spoken in Indonesia?



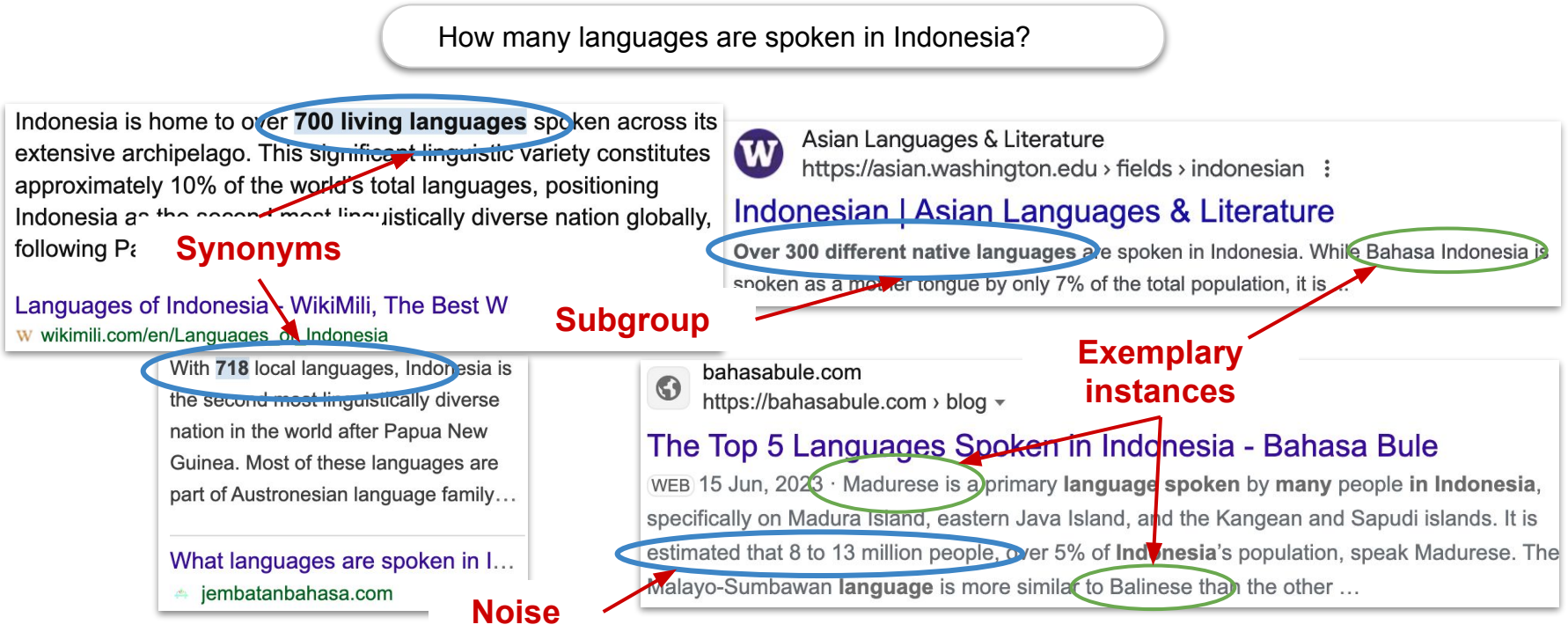
Contributions: Overview

RQ2. How can we aggregate count information from Web snippets?



Contributions: Overview

RQ2. How can we aggregate count information from Web snippets?



Contributions: Overview

RQ1

Counts in KBs (CounQER)

*How to **identify** and **align predicates** that store count information?*

JWS 2020, ESWC 2020 (demo)

RQ2

Counts in Web Snippets (CoQEx)

*How to **aggregate counts** and **enumerations** from Web snippets?*

SIGIR 2022, JoWS 2022 WSDM 2023 (demo)

RQ3

Cardinality Comparison

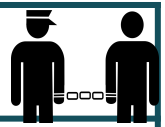
*How to **predict the larger** of two classes of entities?*

WWW 2023

Contributions: Overview

RQ3. How to compare two sets of entities when cardinalities are unreliable?

Are there more **lawyers** than **police officers**?

Sources			
Wikidata	96K	>	8K
SE results	1.3M	>	18K
GPT-3	1.3M	=	1.3M
Ground-truth	3.5M	<	13M

Our approach <

Images credits: https://openclipart.org/
Lawyers and police officers GT: Manually aggregated over countries

Contributions: Overview

RQ1

Counts in KBs (CounQER)

*How to **identify** and **align predicates** that store count information?*

JWS 2020, ESWC 2020 (demo)

RQ2

Counts in Web Snippets (CoQEx)

*How to **aggregate counts** and **enumerations** from Web snippets?*

SIGIR 2022, JoWS 2022 WSDM 2023 (demo)

RQ3

Cardinality Comparison

*How to **predict the larger** of two classes of entities?*

WWW 2023

RQ4

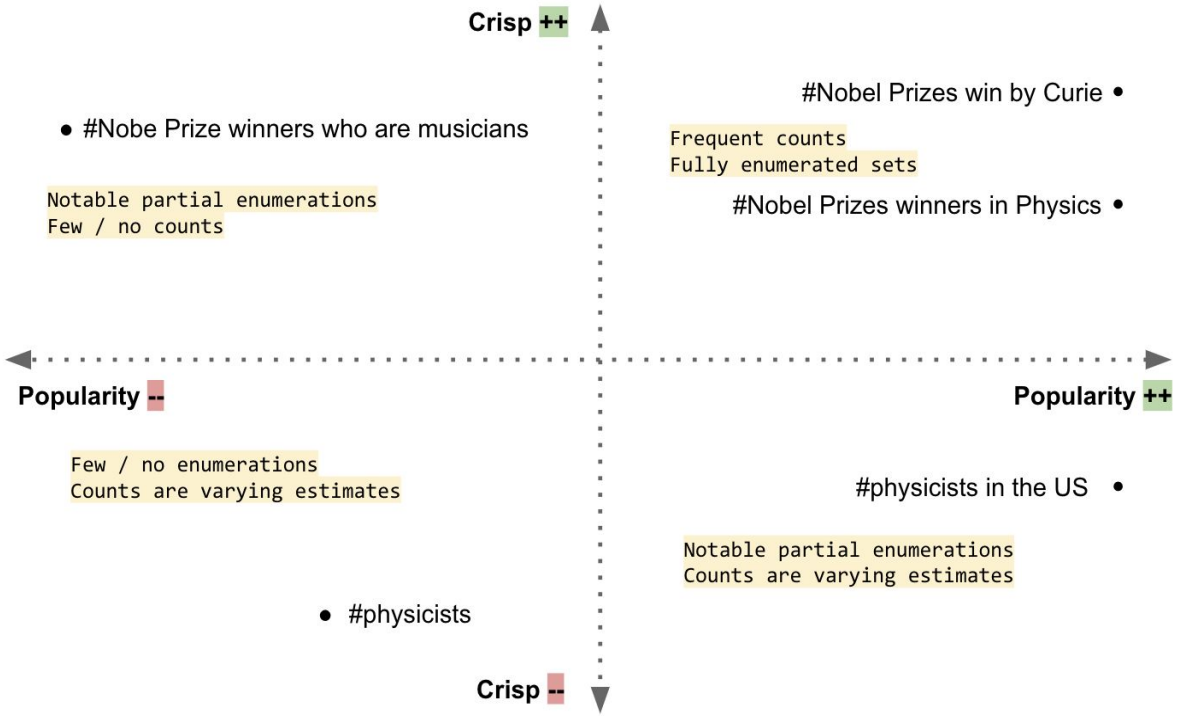
Cardinality Estimation (CardiO)

*How to make **traceable cardinality estimations** in the age of LLMs?*

WWW 2024

Contributions: Overview

RQ4: How to make traceable cardinality estimations in the age of LLMs?



how many physicists are there?

It is difficult to give an exact number of physicists ... However, as of 2021, there are estimated to be over 130,000 physicists globally according to data from the American Physical Society.

Contributions: Overview

RQ1

Counts in KBs (CounQER)

*How to **identify** and **align predicates** that store count information?*

JWS 2020, ESWC 2020 (demo)

RQ2

Counts in Web Snippets (CoQEx)

*How to **aggregate counts** and **enumerations** from Web snippets?*

SIGIR 2022, JoWS 2022 WSDM 2023 (demo)

RQ3

Cardinality Comparison

*How to **predict the larger** of two classes of entities?*

WWW 2023

RQ4

Cardinality Estimation (CardiO)

*How to make **traceable cardinality estimations** in the age of LLMs?*

WWW 2024

2. Contributions

RQ2: Aggregating Count Information from Web Snippets

Contributions: RQ2

How many languages are spoken in Indonesia?

Indonesia is home to over **700 living languages** spoken across its extensive archipelago. This significant linguistic variety constitutes approximately 10% of the world's total languages, positioning Indonesia as the second most linguistically diverse nation globally, following Papua New Guinea.

Languages of Indonesia - WikiMili, The Best Wikipedia Reader
wikimili.com/en/Languages_of_Indonesia

Asian Languages & Literature
https://asian.washington.edu › fields › indonesian

Indonesian | Asian Languages & Literature

Over 300 different native languages are spoken in Indonesia. While Bahasa Indonesia is spoken as a mother tongue by only 7% of the total population, it is ...

With **718** local languages, Indonesia is the second most linguistically diverse nation in the world after Papua New Guinea. Most of these languages are part of Austronesian language family…

What languages are spoken in I…
jembatanbahasa.com

bahasabule.com
https://bahasabule.com › blog

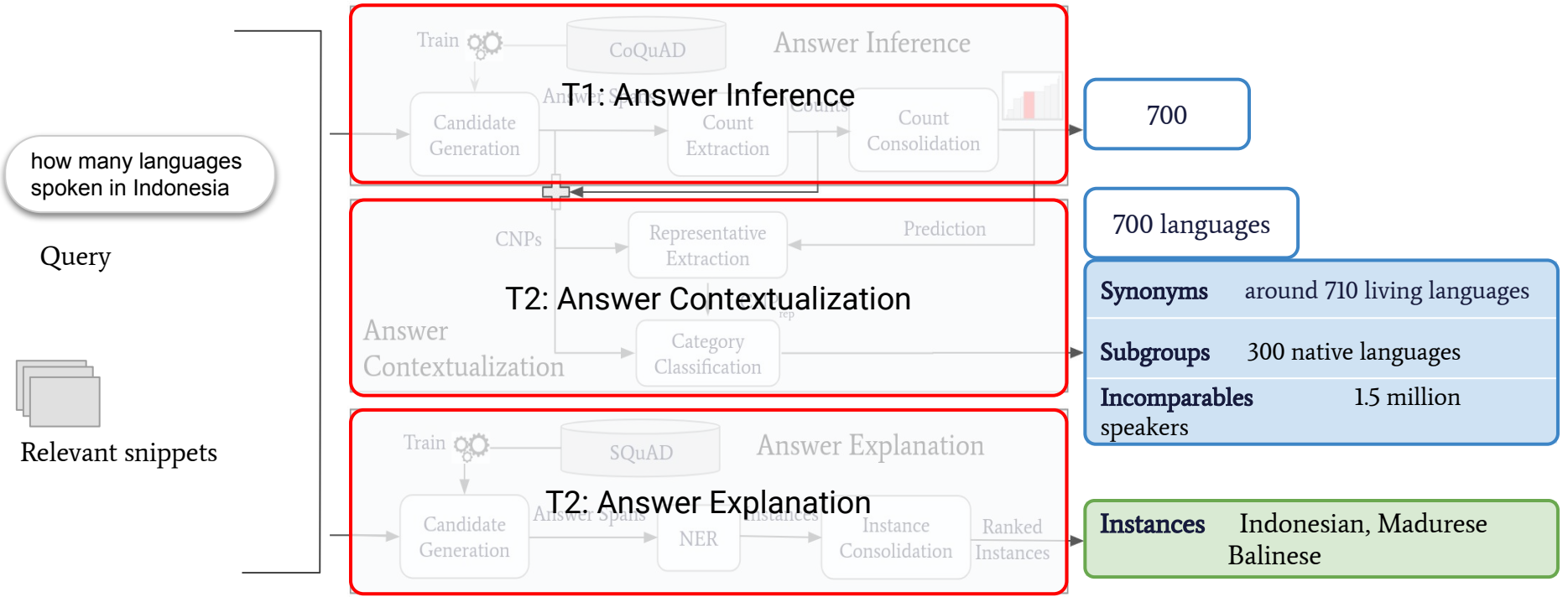
The Top 5 Languages Spoken in Indonesia - Bahasa Bule

WEB 15 Jun, 2023 · Madurese is a primary **language spoken** by **many** people **in Indonesia**, specifically on Madura Island, eastern Java Island, and the Kangean and Sapudi islands. It is estimated that 8 to 13 million people, over 5% of **Indonesia**'s population, speak Madurese. The Malayo-Sumbawan **language** is more similar to Balinese than the other …

Contributions: RQ2: CoQEx

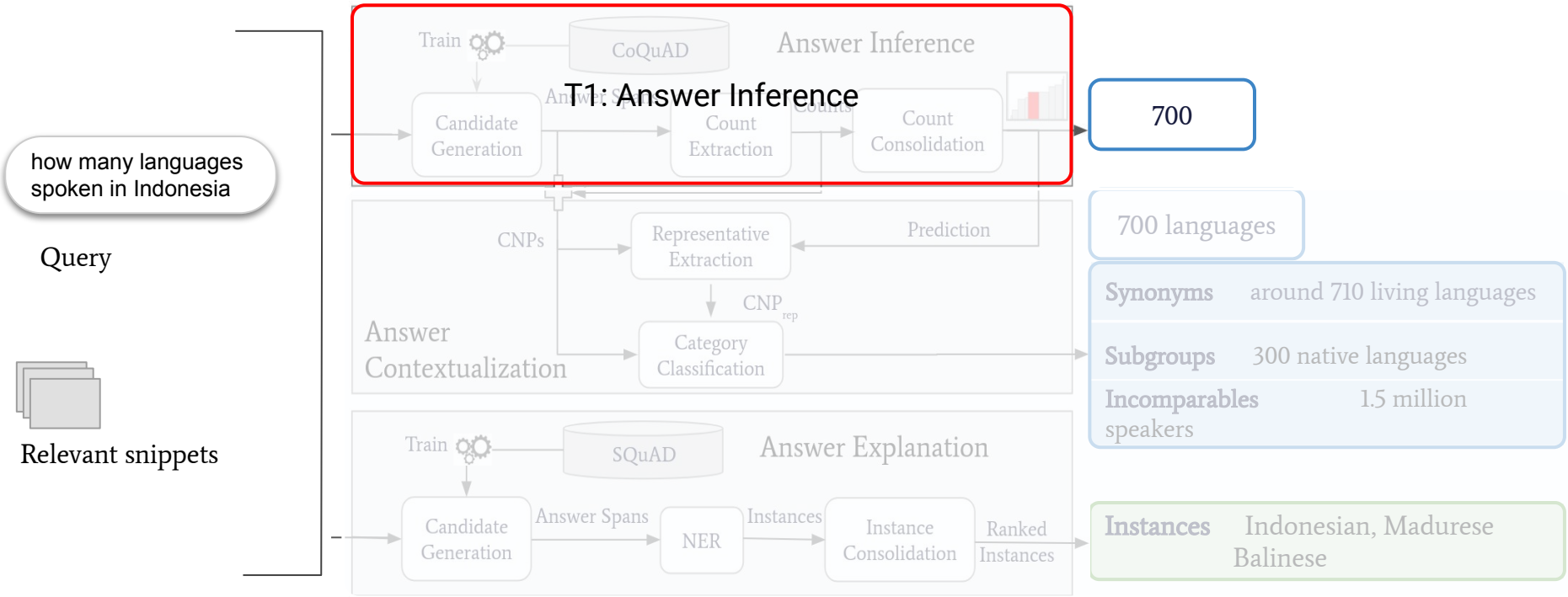
SIGIR 2022, JoWS 2022, WSDM 2023 (demo)

Count **Q**uestion answering with **Ex**planatory evidence



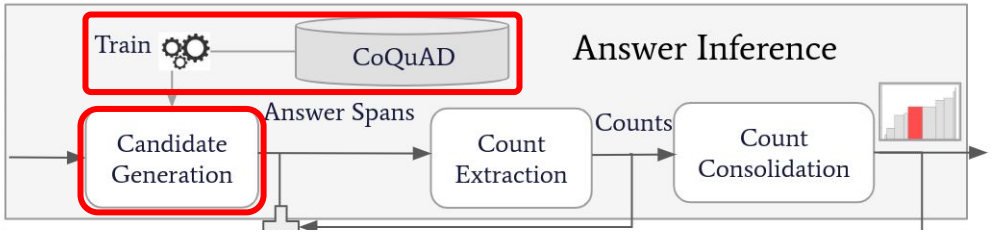
Contributions: RQ2: CoQEx

Count **Q**uestion answering with **Ex**planatory evidence



Contributions: RQ2: Answer Inference

how many languages spoken in Indonesia



Snippet

… There are an estimated 710 living languages spoken in Indonesia

Candidate count contexts (strings)

710 living languages 0.96

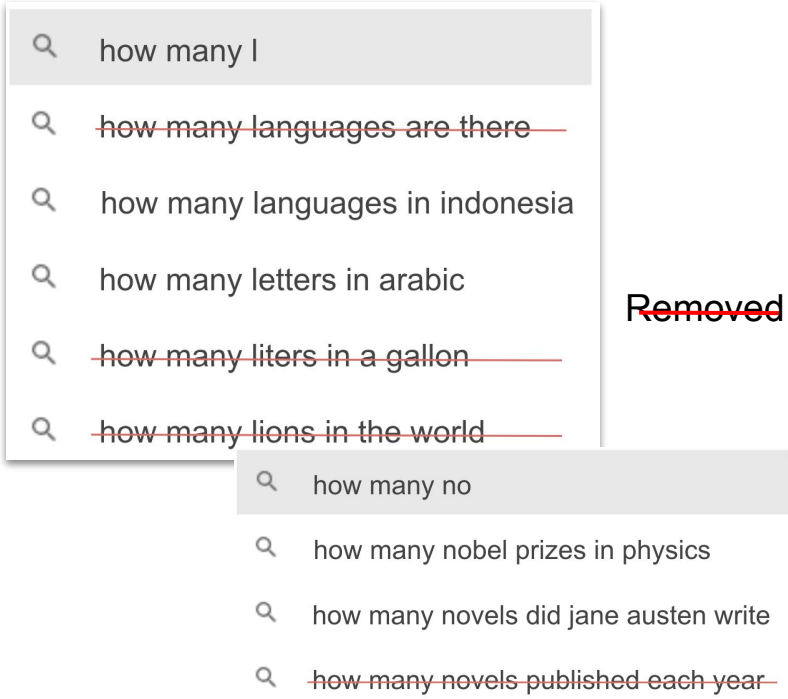
Contributions: RQ2: CoQuAD

Count **Qu**estion **A**nswering **D**ataset

Train a BERT model on candidate extraction.

Automated scraping of count queries.

- Iterative prefixes on Google autocomplete
- Filter out no named-entity, quantity units



Contributions: RQ2: CoQuAD

Count **Qu**estion **A**nswering **D**ataset

Train a BERT model on candidate extraction.

Automated scraping of count queries.

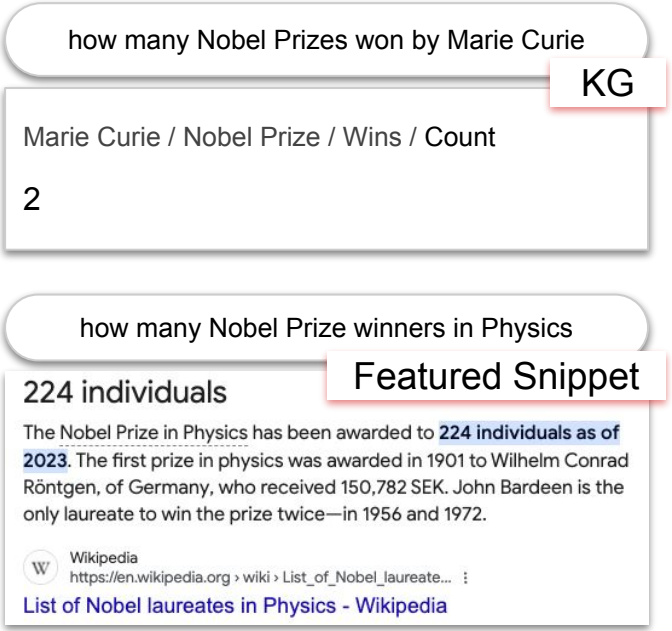
- Iterative prefixes on Google autocomplete
- Filter out no named-entity, quantity units

Automated ground-truth extraction for 5K queries

- KG / featured snippet from Google SERP

Quality evaluation

81% silver answer match with gold answers
84% match within ± 10% of gold answers



Contributions: RQ2: CoQuAD

Count **Qu**estion **A**nswering **D**ataset

Train a BERT model on candidate extraction.

Automated scraping of count queries.

- Iterative prefixes on Google autocomplete
- Filter out no named-entity, quantity units

Automated ground-truth extraction for 5K queries

- KG / featured snippet from Google SERP

Relaxed snippet annotation for 200K web snippets

how many Nobel Prize winners in Physics

224

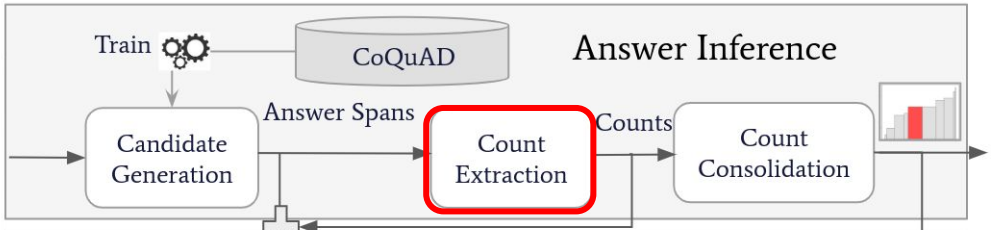
… Physics has been awarded to 224 individuals.

… have been 225 Nobel Prize laureates ..

… Germany has 28 Nobel laureates in Physics..

Contributions: RQ2: Answer Inference

how many languages spoken in Indonesia

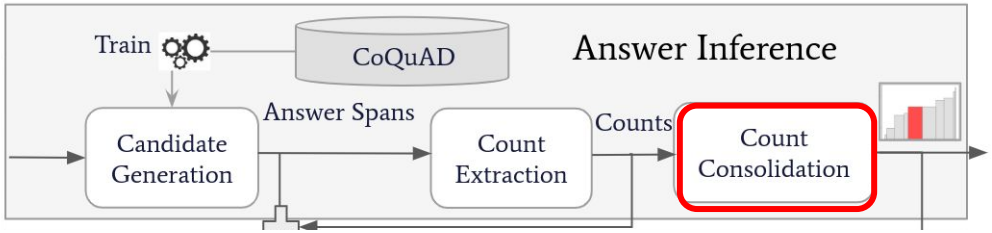


Candidates (strings)		Separate the Counts (integers)
300 native languages	→	300
700 languages	→	700
710 living languages	→	710
1.5 million native speakers	→	1,500,000

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about Quantities in Natural Language. TACL (2015).

Contributions: RQ2: Answer Inference

how many languages spoken in Indonesia



700

Count Distribution

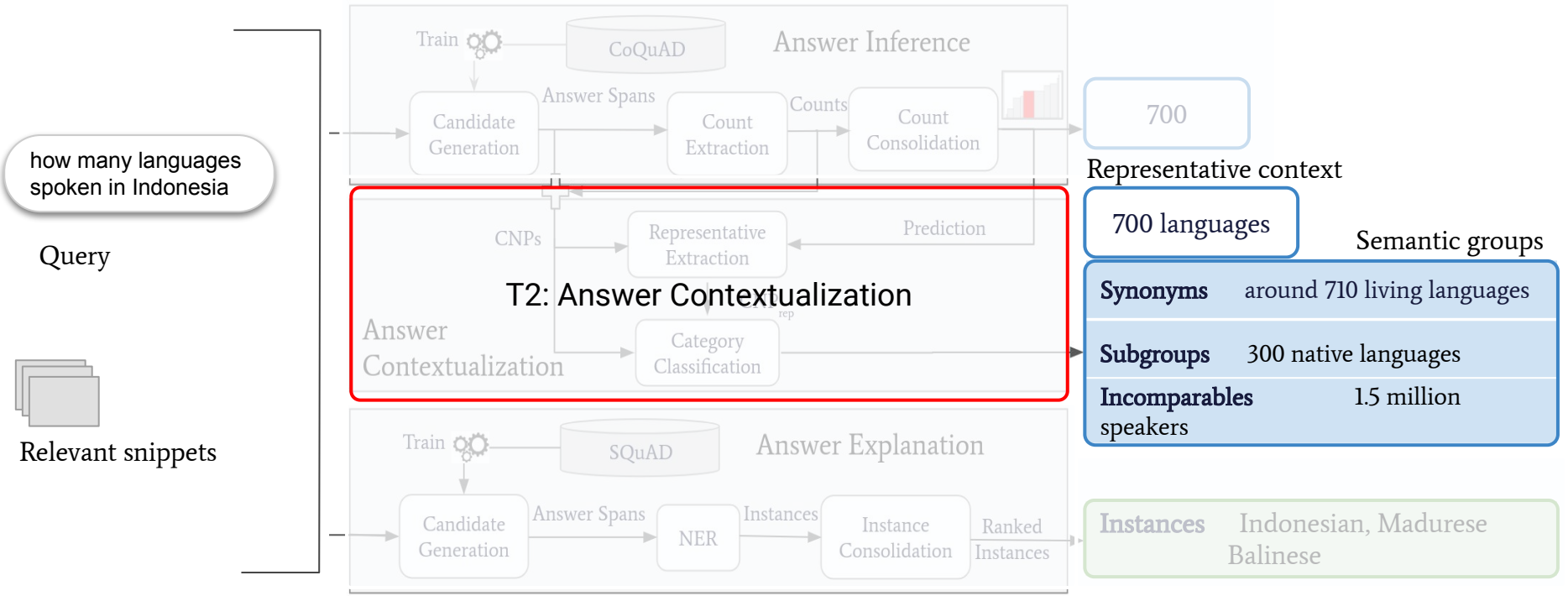
300	0.94	700	0.91	710	0.96	1,500,000	0.61

Strategies

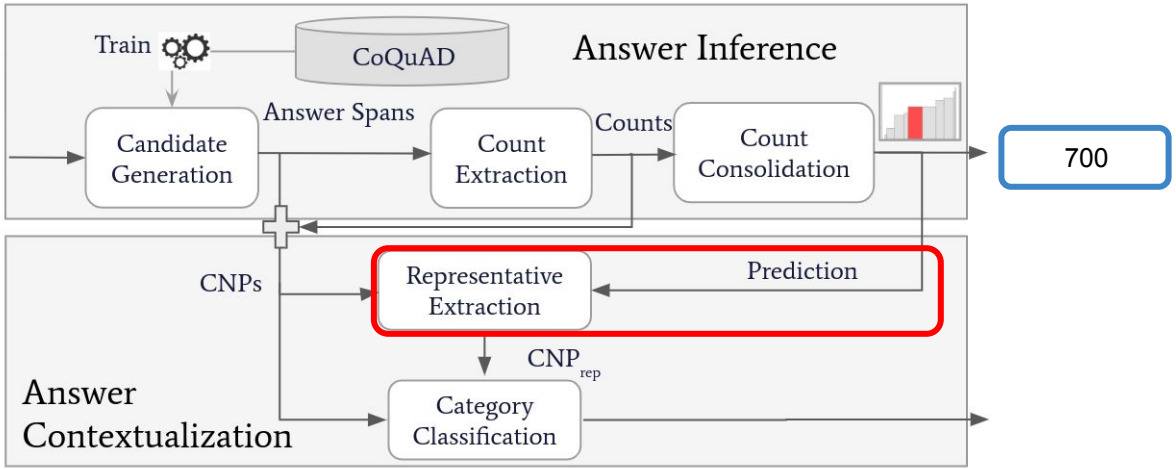
- Median
- Most Frequent
- Most Confident
- Weighted Median

Contributions: RQ2

Count Question answering with Explanatory evidence

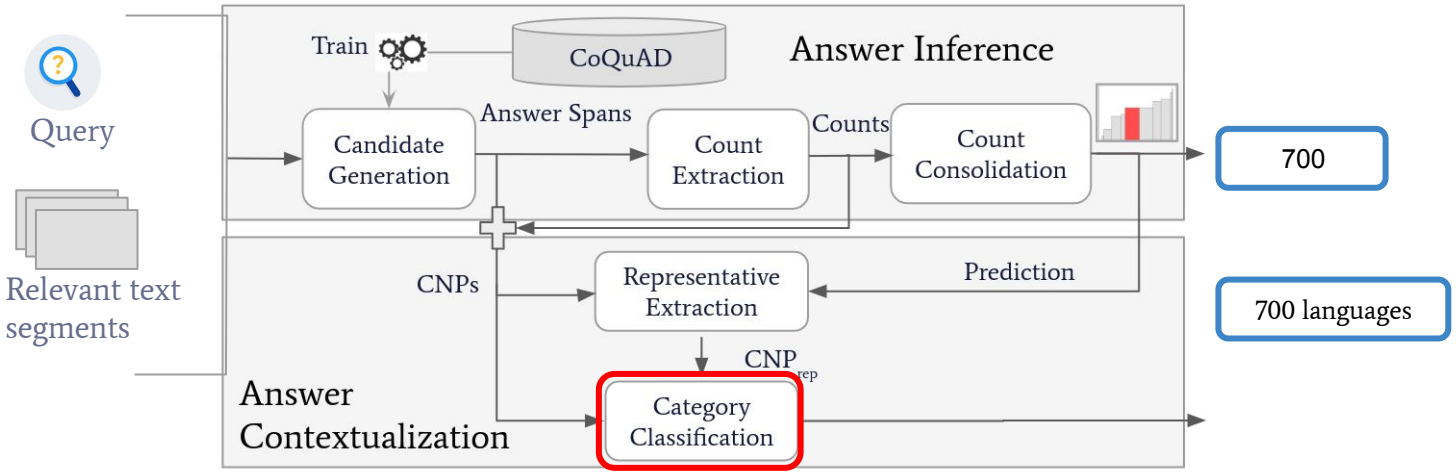


Contributions: RQ2: Answer Contextualization

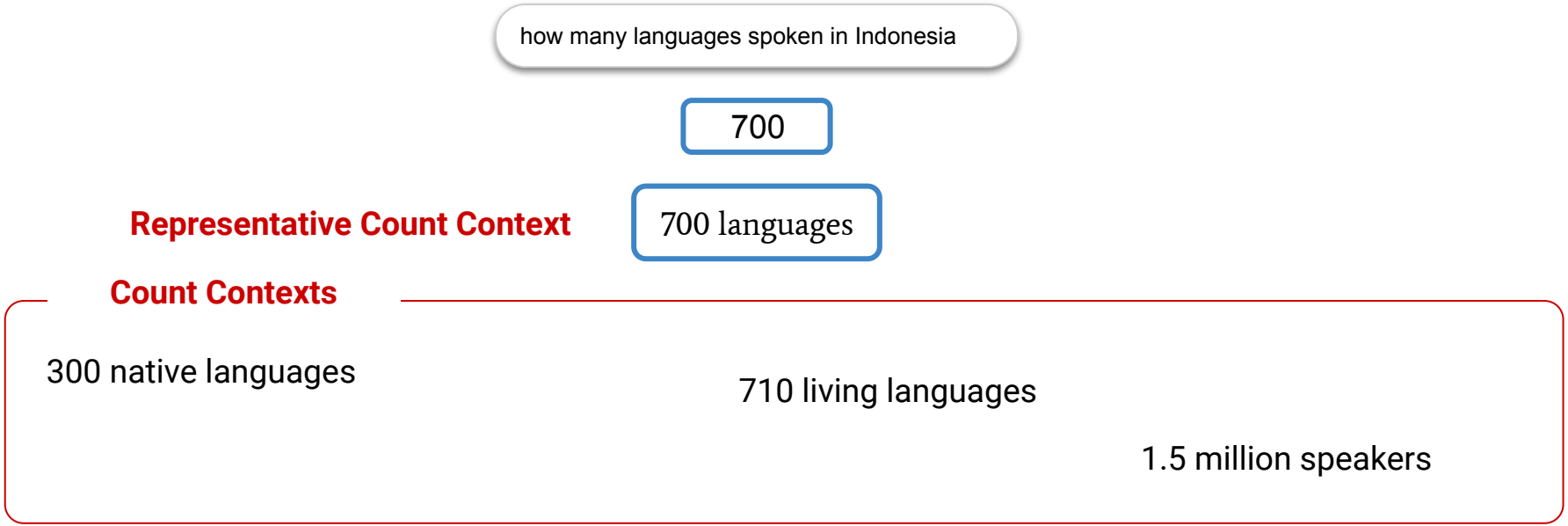


Contributions: RQ2: Answer Contextualization

Contributions: RQ2: Answer Contextualization



Contributions: RQ2: Answer Contextualization



Semantic groups of count contexts w.r.t. the representative count context

- **Numeric distance** of the counts
- **Cosine similarity** of the vector embeddings of count contexts

Contributions: RQ2: Answer Contextualization

how many languages spoken in Indonesia

700

Representative Count Context

700 languages

300 native languages

Subgroups
High cos. similarity
High numeric distance (<)

710 living languages

Synonyms
High cos. similarity
Low numeric distance

1.5 million speakers

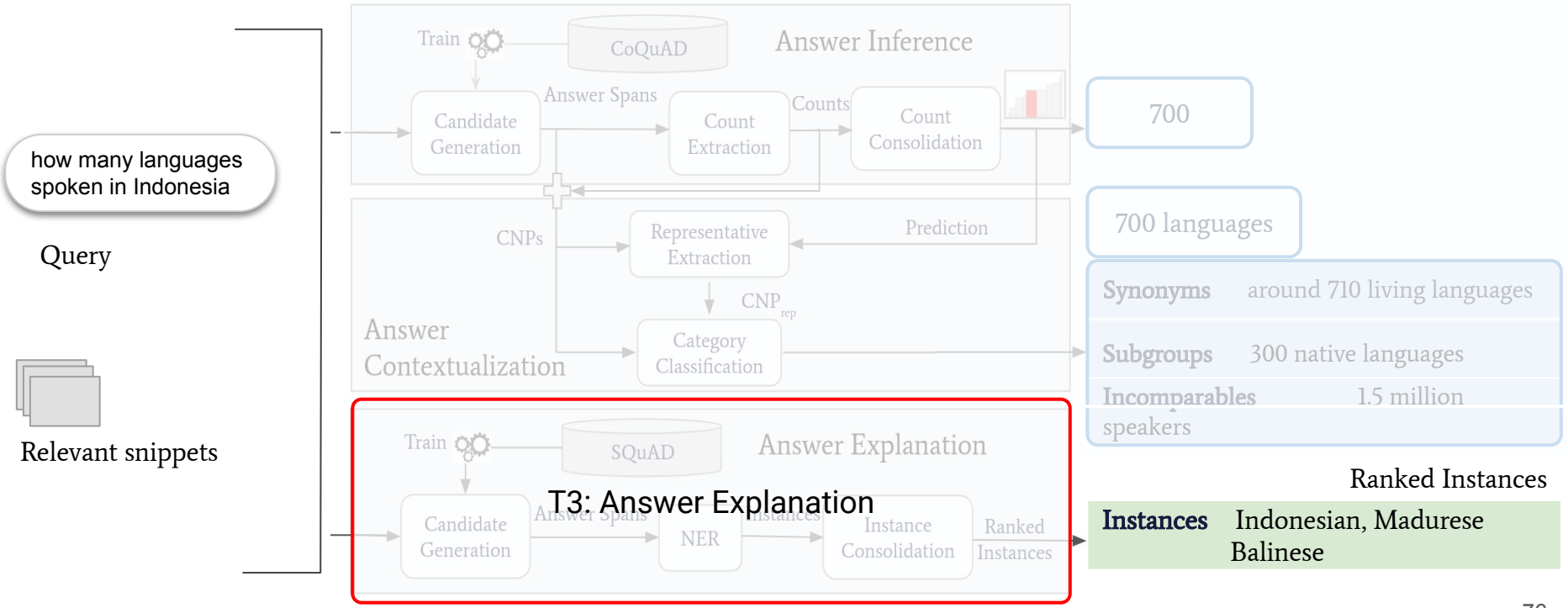
Incomparables
Low cos. similarity
or
High cos. similarity
High numeric distance (>)

Semantic groups of count contexts w.r.t. the representative count co

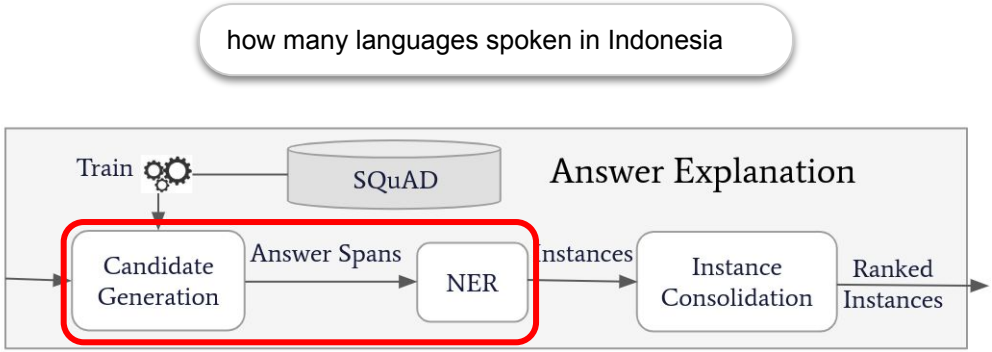
- **Numeric distance** of the counts
- **Cosine similarity** of the vector embeddings of count contexts

Contributions: RQ2

Count Question answering with Explanatory evidence



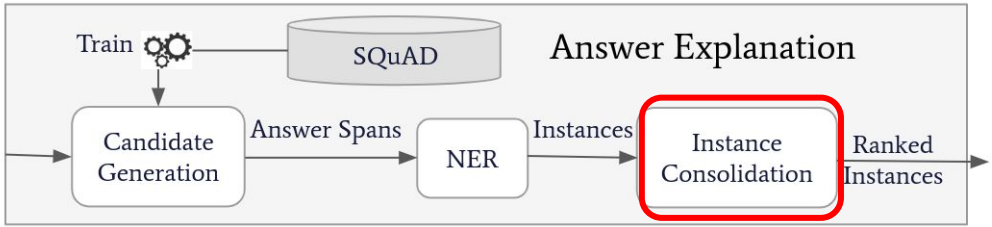
Contributions: RQ2: Answer Explanation



Snippets (Candidate Answer Spans)	→	**Instances (named-entities)**
Madurese and Indonesian are spoken, ..		Madurese
		Indonesian
Other commonly used languages are Balinese, ..		Balinese

Contributions: RQ2: Answer Explanation

how many languages spoken in Indonesia



Instances
Madurese
Indonesian
Balinese

Madurese 0.97

Indonesian 0.97

Balinese 0.78

Strategies

- No consolidation
- Context Frequency
- Summed Confidence
- Type Compatibility

Contributions: RQ2: CoQEx

nlcounquer.mpi-inf.mpg.de

Demo

Contributions: RQ2: Results

Baselines

QAnswer [Diefenbach et al. (2019)] QA over KB

GoogleSDA QA over the Web via search-engine direct answer

Contributions: RQ2: Results

Baselines

QAnswer [Diefenbach et al. (2019)] QA over KB

GoogleSDA QA over the Web via search-engine direct answer

Metrics (Answer Inference)

Relaxed Precision (RP) answer is within ± α of GT

Coverage (Cov) fraction of non-empty answers

P/C RP vs Cov tradeoff

Metrics (Answer Explanation)

Mean Average Precision (MAP) fraction of relevant instances

Average Recall (AR) fraction of ground-truth entities returned

Hit@k fraction of queries with ≥ 1 relevant instances in the top-k

Mean Reciprocal Rank (MRR) inverse rank of the 1st relevant instance

Contributions: RQ2: Results

Answer Inference

CoQEx provides better trade-off between relaxed precision and coverage.

System	CoQuAD			LCQuAD_{count}			Stresstest			NaturalQuestions		
	RP	Cov	P/C	RP	Cov	P/C	RP	Cov	P/C	RP	Cov	P/C
QAnswer [Diefenbach et al.(2019)]	6.6	96.2	12.4	45.0	96.1	**61.3**	9.0	100	16.5	12.5	98.8	22.1
GoogleSDA	93.2	18.3	30.6	44.4	8.6	14.4	79.3	29.0	42.4	94.4	22.6	36.4
CoQEx	37.7	84.7	**52.2**	13.6	49.3	21.3	43.6	91.6	**59.1**	43.0	91.6	**58.5**

System	KG (50)			Snippet (172)			NDA (100)		
	RP	Cov	P/C	RP	Cov	P/C	RP	Cov	P/C
QAnswer [Diefenbach et al.(2019)]	12.2	98.0	21.7	4.1	97.0	8.0	8.2	94.0	15.1
GoogleSDA	100	100	**100**	75.0	2.3	4.5	40.0	5.0	8.8
CoQEx	23.1	98.0	37.4	45.3	85.8	**59.3**	31.9	76.3	**45.0**

by CoQuAD GT source

CoQEx shines in queries with no direct answers or queries answerable by featured snippets.

Contributions: RQ2: Results

Answer Explanations

Baselines perform well on KG-answerable queries
GoogleSDA returns **no instances** for snippet-answerable queries.

System	MAP@1	MAP@5	MAP@10	AR@1	AR@5	AR@10	Hit@10	MRR
CoQuAD (142 queries)								
QAnswer [Diefenbach et al.(2019)]	8.5	9.3	9.6	2.9	6.5	8.4	19.7	0.118
GoogleSDA	**14.8**	**12.8**	**10.6**	**4.8**	**13.7**	**14.3**	23.2	0.185
CoQEx	12.0	11.7	11.0	2.3	9.3	12.7	**37.3**	**0.200**
KG (50 queries)								
QAnswer [Diefenbach et al.(2019)]	20.0	21.3	22.0	6.1	14.2	18.5	38.0	0.250
GoogleSDA	**42.0**	**36.4**	**30.0**	**13.5**	**38.9**	**40.7**	**66.0**	**0.526**
CoQEx	14.0	13.7	12.9	3.8	13.0	18.4	42.0	0.233
Snippet (92 queries)								
QAnswer [Diefenbach et al.(2019)]	2.2	2.8	2.9	1.2	2.4	3.0	9.8	0.046
GoogleSDA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CoQEx	**10.9**	**10.6**	**10.0**	**1.5**	**7.3**	**9.7**	**34.8**	**0.182**

by CoQuAD GT source

CoQEx shines in snippet-answerable queries.

Contributions: RQ2: User Study

Annotator Precision on Answer Inference

Class	Only Count	+Instances	+CNPs	+Snippet	All
Correct	73	63	78	75	**88**
Incorrect	28	45	40	**53**	45
Both	55	56	63	66	**71**

Annotator precision (in %age)

User can better distinguish between correct and incorrect system answers, especially incorrect, with the help of different forms of CoQEx explanations.

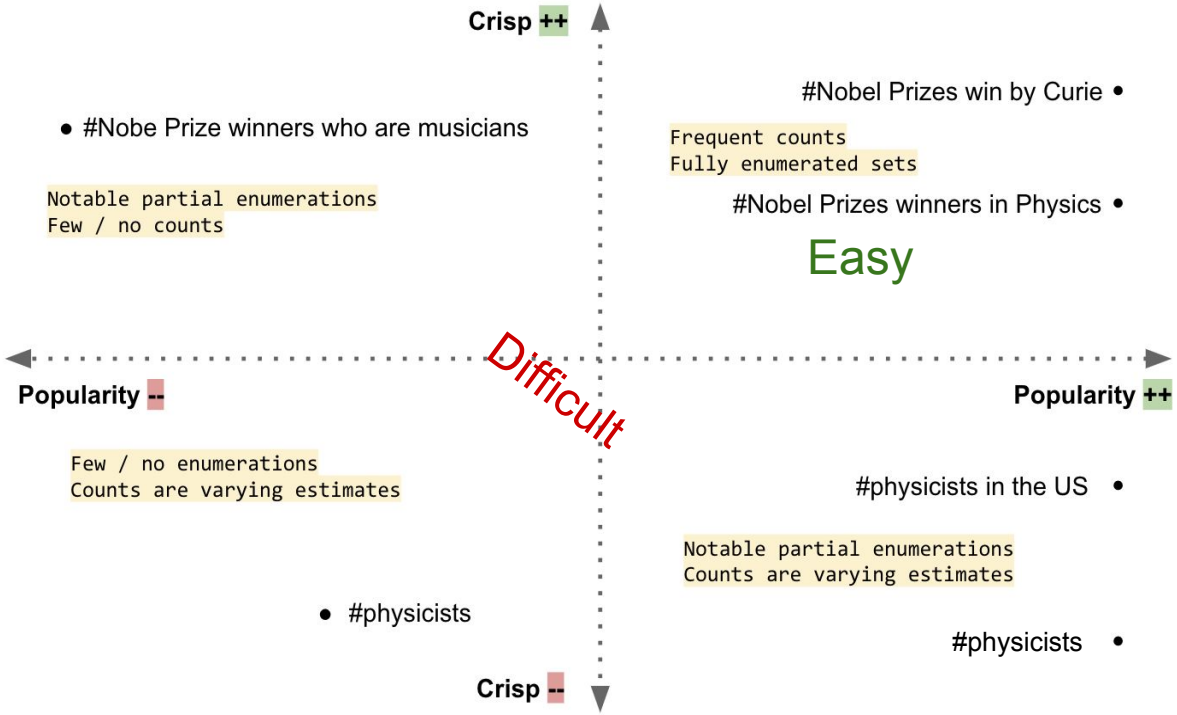
2. Contributions

RQ4: Cardinality Estimation

GPT4-o and DALL-E AI

Contributions: RQ4

RQ4: How to make traceable cardinality estimations in the age of LLMs?



how many physicists are there?

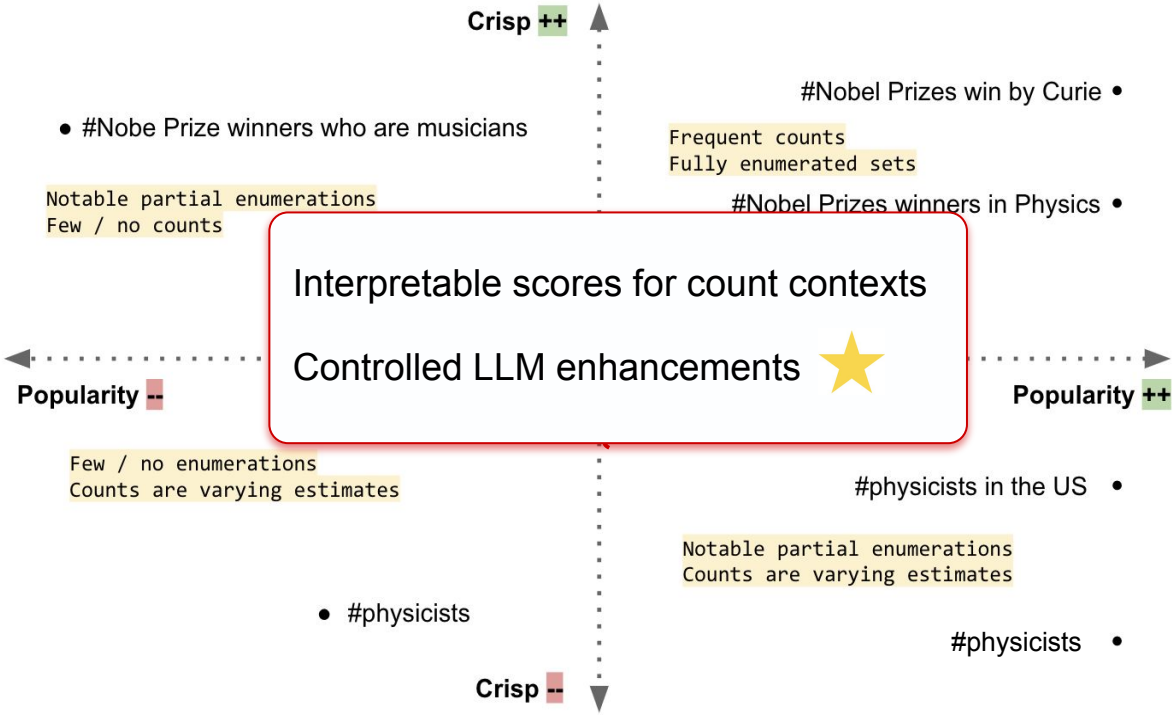
It is difficult to give an exact number of physicists ... However, as of 2021, there are estimated to be over 130,000 physicists globally according to data from the American Physical Society.

how many physicists are there?

It is difficult to provide an exact number of physicists worldwide as the number is constantly changing due to new individuals entering the field and others leaving. However, according to the American Physical Society, there are over 50,000 members in their organization alone. When considering physicists globally, the number is likely in the hundreds of thousands.

Contributions: RQ4

RQ4: How to make traceable cardinality estimations in the age of LLMs?



how many physicists are there?

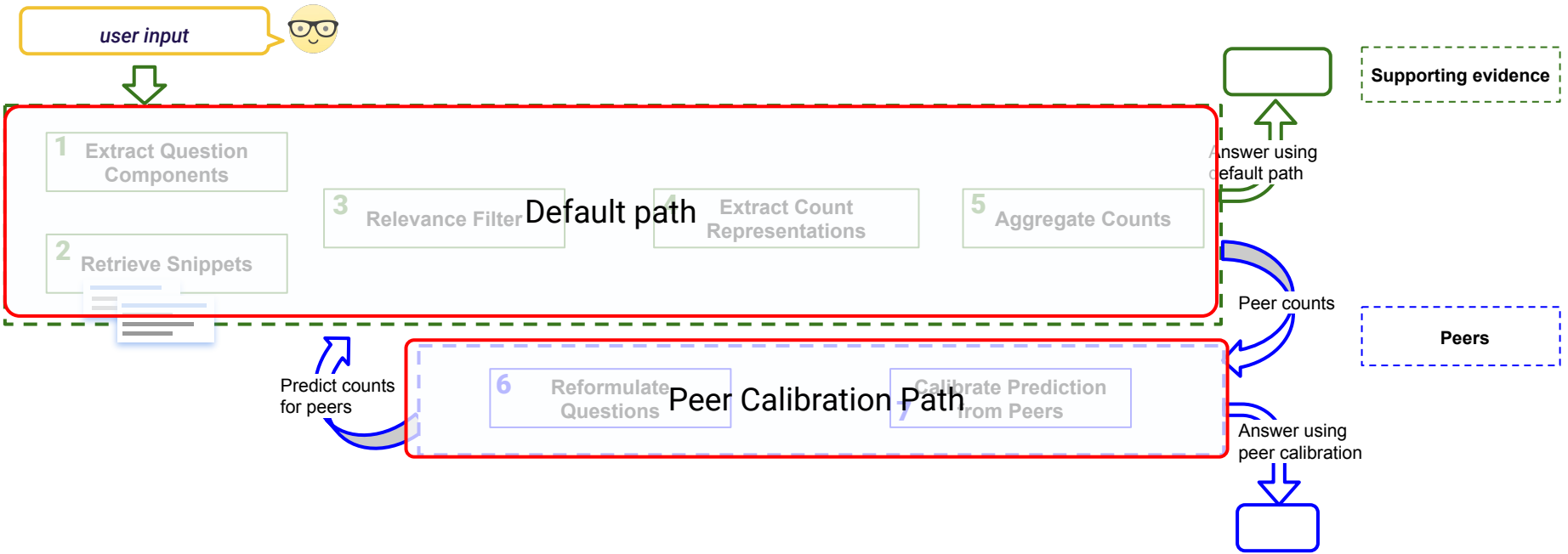
It is difficult to give an exact number of physicists ... However, as of 2021, there are estimated to be over 130,000 physicists globally according to data from the American Physical Society.

how many physicists are there?

It is difficult to provide an exact number of physicists worldwide as the number is constantly changing due to new individuals entering the field and others leaving. However, according to the American Physical Society, there are over 50,000 members in their organization alone. When considering physicists globally, the number is likely in the hundreds of thousands.

Contributions: RQ4: CardiO *WWW 2024*

Cardinality Predictor from **O**nline Sources



Contributions: RQ4: Default Path

Sentence-level relevance of counts

Cos. Sim. question and snippet	
.. 17,000 Artificial Intelligence companies	0.85
.. ~~startups in 46 US states ..~~	0.43
.. startups .. approx. 14,000 .. in AI in US..	0.92
.. 524 startups founded in US in 2022.	0.69

LLM Filter
✓ .. 17,000 Artificial Intelligence companies
~~.. startups in 46 US states ..~~
✓ .. startups .. approx. 14,000 .. in AI in US..
✓ .. 524 startups founded in US in 2022.

Contributions: RQ4: Count Representation

Track context of count in three levels: type, sentence, snippet

4 Extract Count Representations

5 Aggregate Counts

Answer using default path

Supporting evidence

Extract with using dependency parsers or LLMs

Sentence	count	type	sentence	snippet
.. 17,000 Artificial Intelligence companies	17,000	Artificial Intelligence companies		
.. startups .. approx. 14,000 .. in AI in US..	14,000			
.. 524 startups founded in US in 2022.	524	startups		

Contributions: RQ4: Count Aggregation

Relevance and consistency

Relevance of source (sentence, snippet) to question
Relevance of context type to answer type (*AI startups*)
and
nearest-neighbour consistency

5 Aggregate Counts

17,000 Artificial Intelligence companies

startups .. 14,000 .. in AI

Supporting evidence

Count Representation (c)

count	type	sentence	snippet
17,000	Artificial Intelligence companies		
14,000			
524	startups		

Consistency(c)

$$\beta R(c) + (1 - \beta) R_{cons}(c)$$

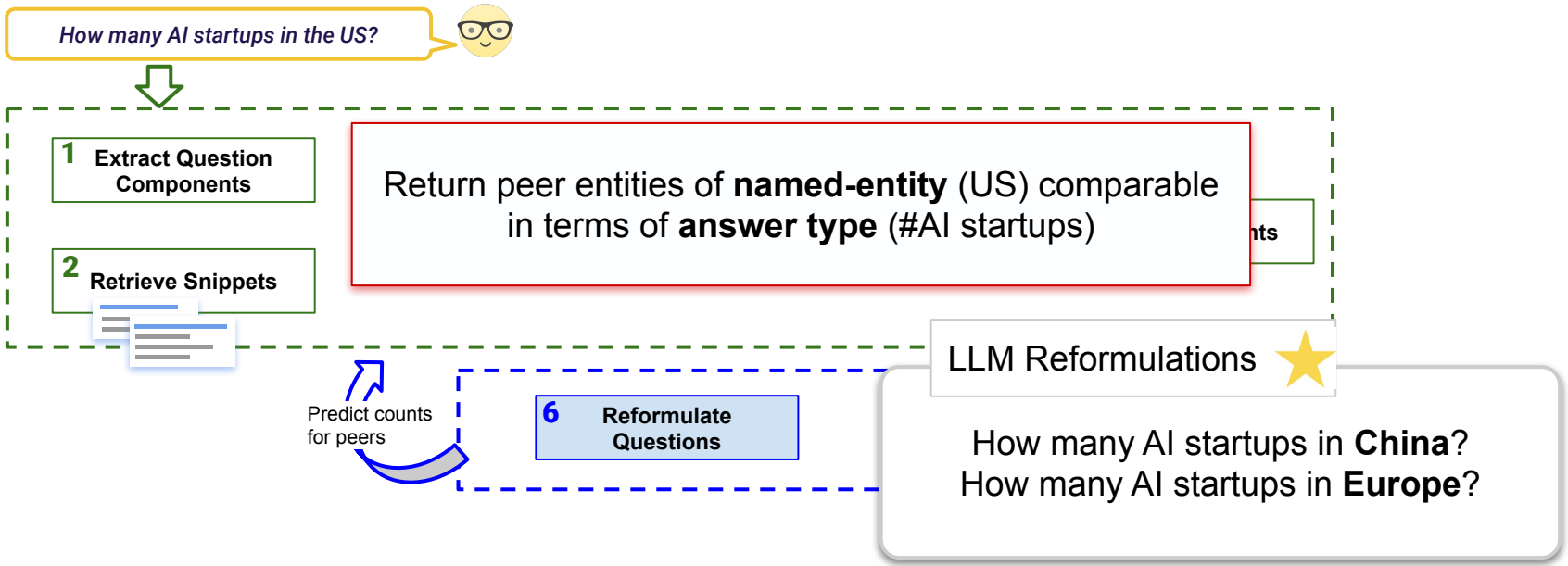
524 startups

dropped

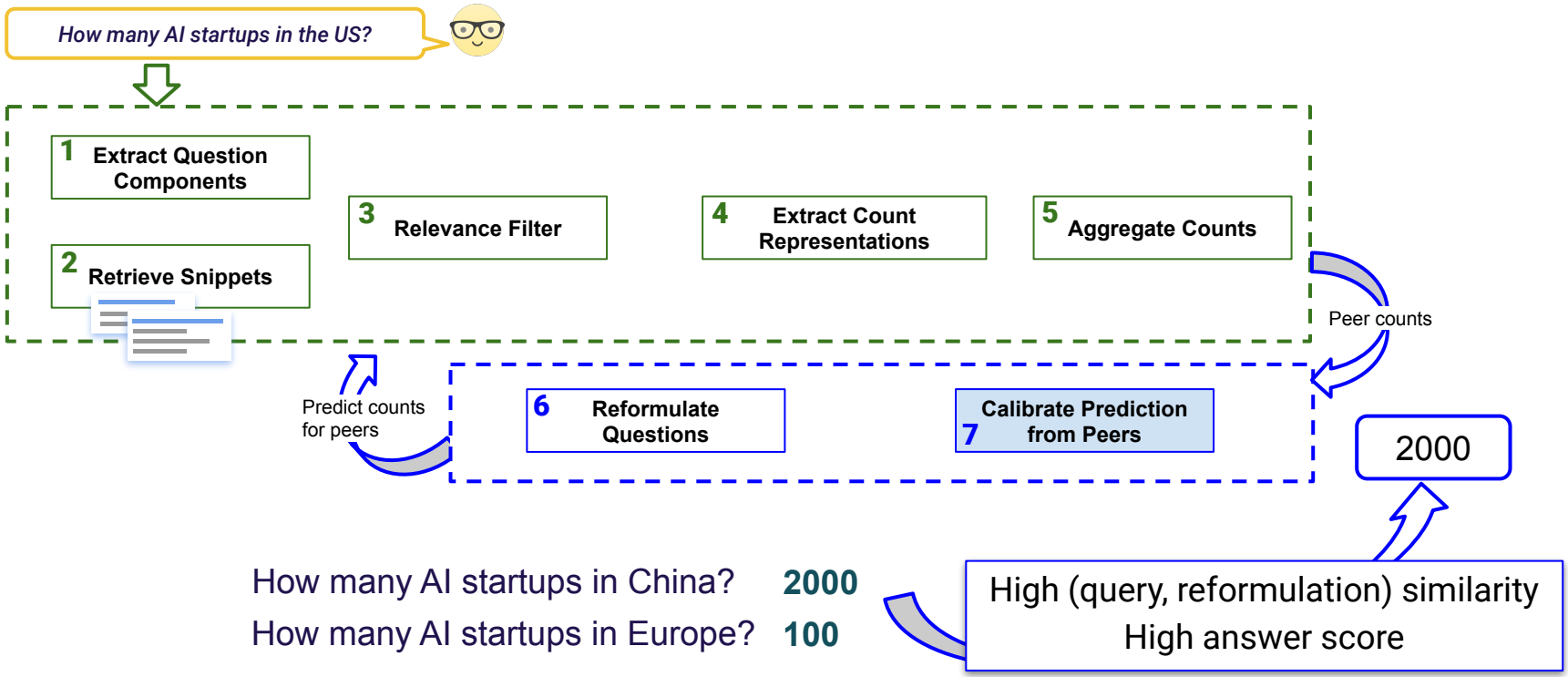
Relevance score of the count repr.

Combined relevance scores of k-nearest count repr.

Contributions: CardiO: Peer Calibration Path



Contributions: CardiO: Peer Calibration Path



Contributions: RQ4: Results

Metrics

Order-of-magnitude precision and recall (OMP, OMR)

Baselines

CoQEx

Llama2

GPT3.5

Contributions: RQ4: Results

Metrics

Order-of-magnitude precision and recall (OMP, OMR)

Baselines

CoQEx	Fully traceable, Model parameters ~100M	
Llama2	7B, 70B	Pretrained, Finetuned
GPT3.5	~100B	Finetuned
	Model parameters	Model Training

Contributions: RQ4: Results

Metrics

Order-of-magnitude precision and recall (OMP, OMR)

Baselines

CoQEx	Fully traceable, Model parameters ~100M	
Llama2	0-shot:	None
GPT3.5	Snippet-augmented:	Limited
	Traceability	

Contributions: RQ4: Results

Benchmarks

Cardinality Questions (CQ): 500 high-quality annotated questions

Question characteristics: popularity, specificity, presence of named-entity
Ground-truth characteristics: exact/estimate, direct/aggregated, time variance

CoQuAD: 312 questions

Natural Questions (NQ): 84 questions

Contributions: RQ4: Results

CardiO vs. LLM

- CardiO at par / better than Llama models in 0-shot setting.
- CardiO ahead of snippet-augmented Llama models in recall.

Answer Traceability	Method	CQ (n=500)			NQ (n=84)			CoQuAD (n=312)		
		EP	OMP	OMR	EP	OMP	OMR	EP	OMP	OMR
Not possible	0-shot LLAMA2-7B-chat	0.056	0.578	0.560	0.250	0.695	0.662	0.137	0.584	0.547
	0-shot LLAMA2-70B	0.113	0.670	0.521	0.288	0.725	0.630	0.190	0.692	0.548
	0-shot LLAMA2-70B-chat	0.079	0.653	0.631	0.325	0.775	0.738	0.220	0.646	0.621
	0-shot GPT3.5	0.133	0.716	0.689	0.438	0.797	0.759	0.273	0.724	0.689
Medium	Snippets + LLAMA2-7B-chat	0.213	0.640	0.553	0.289	0.674	0.610	0.224	0.635	0.572
	Snippets + LLAMA2-70B	0.242	0.723	0.567	0.338	0.723	0.611	0.320	0.743	0.581
	Snippets + LLAMA2-70B-chat	0.278	0.743	0.728	0.494	0.838	0.808	0.302	0.749	0.715
	Snippets + GPT3.5	0.303	0.825	0.751	0.548	0.834	0.724	0.381	0.815	0.726
High	CoQEx [Ghosh et al.(2022a)]	0.175	0.631	0.577	0.329	0.665	0.626	0.266	0.696	0.611
	CardiO (vanilla)	0.192	0.659	0.659	0.298	0.661	0.661	0.228	0.618	0.618

Contributions: RQ4: Results

CardiO vs. LLM

- CardiO at par / better than Llama models in 0-shot setting.
- CardiO ahead of snippet-augmented Llama models in recall.
- Questions where CardiO outperforms Llama w. snippets:
  - no named-entity
  - fuzzy entity set

Contributions: RQ4: Results

CardiO vs. LLM

- CardiO at par / better than Llama models in 0-shot setting.
- CardiO ahead of snip
- Questions where Car
  - no named-entity
  - fuzzy entity set

how many AI researchers are there

Llama2-70B w. snippets

Ground truth: 300,000

Generated text: *"According to the article, there are approximately 300 full-time technical AI safety researchers, 100 full-time non-technical AI safety researchers, and 400 AI safety researchers in total today. ..(one more explanation sentence)"*

Extracted counts: [300, 100, 400] (we evaluate the first count)

CardiO

CardiO prediction: 300,000

Source context: just 300,000 AI researchers

Source snippet: blueTencent says there are only 300,000 AI engineers worldwide, but ... According to the study, compiled by the Tencent Research Institute, there are just 300,000 "AI researchers and practitioners" worldwide, but the "market demand" is for millions of roles.

Contributions: RQ4: Results

LLM enhancements in CardiO

- Sentence filter and peer calibration provide minimal improvements
- Count extraction decreases overall performance

Method	CQ (n=500)			NQ (n=84)			CoQuAD (n=312)		
	EP	OMP	OMR	EP	OMP	OMR	EP	OMP	OMR
CoQEx [Ghosh et al.(2022a)]	0.175	0.631	0.577	0.329	0.665	0.626	0.266	0.696	0.611
CardiO (vanilla)	0.192	0.659	0.659	0.298	0.661	0.661	0.228	0.618	0.618
+ LLM Sentence filter	0.191	0.663	0.653	0.293	0.657	0.634	0.229	0.657	0.634
+ LLM Count extraction	0.174	0.571	0.571	0.298	0.661	0.661	0.205	0.601	0.601
+ Peer calibration	0.190	0.657	0.657	0.310	0.654	0.654	0.234	0.623	0.623

Contributions: RQ4: Results

Aggregation based on neighbours

- No significant performance improvement over only relevance
- No performance degradation → serve as reliable supporting evidence

Method	CQ (n=500)			NQ (n=84)			CoQuAD (n=312)		
	EP	OMP	OMR	EP	OMP	OMR	EP	OMP	OMR
CoQEx [Ghosh et al.(2022a)]	0.175	0.631	0.577	0.329	0.665	0.626	0.266	0.696	0.611
CardiO (vanilla)	0.192	0.659	0.659	0.298	0.661	0.661	0.228	0.618	0.618
+ Consistent aggregation	0.186 (±0.064)	0.653 (±0.023)	0.653 (±0.023)	0.274	0.660	0.660	0.240	0.642	0.642

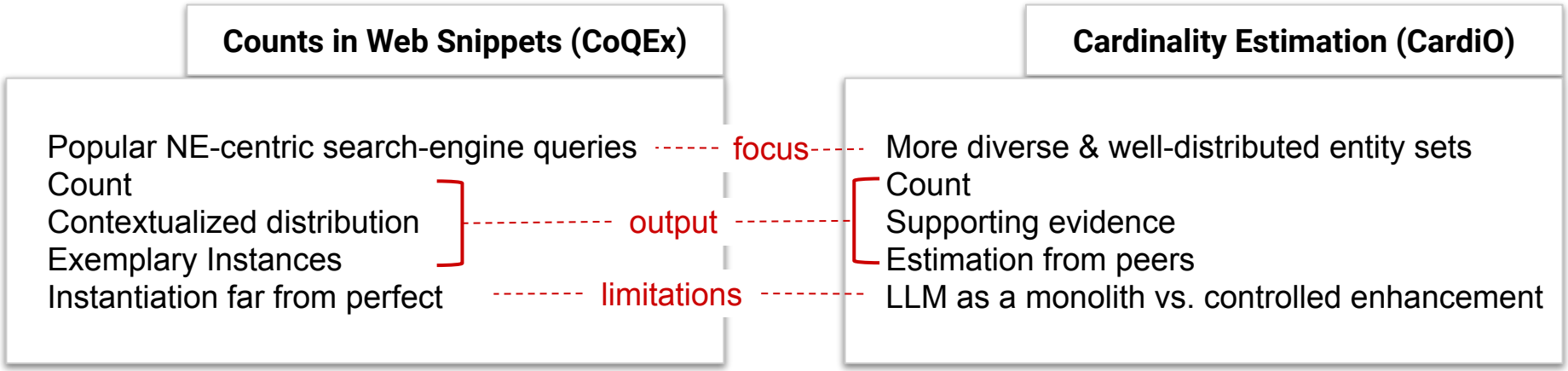
3. Summary

Summary

- Count information can expressed as **a cardinality** or **enumerations**
- Knowledge on the Web is in **explicit** and **latent** format

Summary

- Count information can expressed as **a cardinality** or **enumerations**
- Knowledge on the Web is in **explicit** and **latent** format



4. Limitations & Future Work

Limitations

* Training and evaluation data
    - Crowd-sourced dataset → quality control
    - Semi-automated dataset → popularity bias
    - Expert annotated dataset → scalability issues

Limitations

- Training and evaluation data
  - Crowd-sourced dataset → quality control
  - Semi-automated dataset → popularity bias
  - Expert annotated dataset → scalability issues
- Predicting strict bounds is hard
  - Mathematical relations (<,>,~,=) extracted but unused
  - Semantic grouping could be principled

how many AI startups?

more than 1000 startups upper bound?
500 GenAI companies in US in 2020 lower bound?

Future Work

- Numbers are yet more tokens for LLMs

  Supervised training to extract structured count representations.

  Counting capability is underexplored

- Reasoning in LLMs is not well-understood

  Focus on explainability, traceability, verification.

- Count information as exploratory search

  Popular subgroups / peers to estimate bounds.

- Beyond counts of named-entities

  Numbers in sports / financial analysis.

5. Conclusion

Conclusion

Counts in KBs (CounQER)

counquer.mpi-inf.mpg.de/spo

Uncovering Hidden Semantics of Set Information in Knowledge Bases
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *Journal of Web Semantics 2020.*

CounQER: A System for Discovering and Linking Count Information in Knowledge Bases
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *System demonstration at ESWC 2020.*

Counts in Web Snippets (CoQEx)

nlcounquer.mpi-inf.mpg.de

Answering Count Queries with Explanatory Evidence
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *SIGIR 2022.*

CoQuAD

Answering Count Questions with Structured Answers from Text
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *Journal of Web Semantics 2023.*

CoQEx: Entity Counts Explained
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. System Demonstration at WSDM 2023.

Cardinality Comparison

Class Cardinality Comparison as a Fermi Problem
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *Web Conference 2023.*

Cardinality Estimation (CardiO)

CardiO: Predicting Cardinality from Online Sources
Shrestha Ghosh, Simon Razniewski, Damien Graux, Gerhard Weikum. *Web Conference 2024.*

CQ

Conclusion

Counts in KBs (CounQER)

counquer.mpi-inf.mpg.de/spo

Uncovering Hidden Semantics of Set Information in Knowledge Bases
Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *Journal of Web Semantics 2020.*

CounQER: A System for Discovering and Linking Count Information in Knowledge Bases
onstration at ESWC 2020.

Counts in Web Snippets (CoQEx)

nlcounquer.mpi-inf.mpg.de

CoQuAD

eb Semantics 2023.

onstration at WSDM 2023.

Cardinality Comparison

Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. *Web Conference 2023.*

Cardinality Estimation (CardiO)

CQ

CardiO: Predicting Cardinality from Online Sources
Shrestha Ghosh, Simon Razniewski, Damien Graux, Gerhard Weikum. *Web Conference 2024.*

Bibliography

[Suchaneck et al. (2007)] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (WWW 2007).

[Auer et al. (2007)] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In International Semantic Web Conference.

[Vrandečić (2012)] Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In Proceedings of the 21st international conference on World Wide Web (WWW 2012).

[Singhal (2012)] Amit Singhal. 2012. Introducing the knowledge graph: Things, not strings. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

[Saha et al. (2017)] Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for Numerical Open IE. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2017).

[Ho et al. (2022)] Vinh Thinh Ho, Daria Stepanova, Dragan Milchevski, Jannik Strötgen, and Gerhard Weikum. 2022. Enhancing knowledge bases with quantity facts. In Proceedings of the Web Conference 2022 (WWW 2022).

[Liu and Singh (2004)] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal (2004).

[Jain et al. (2020)] Prachi Jain, Sushant Rathi, Soumen Chakrabarti, et al. 2020. Temporal Knowledge Base Completion: New Algorithms and Evaluation Protocols. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).

[Arnaout et al. (2021)] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z Pan. 2021. Negative statements considered useful. Journal of Web Semantics (2021).

Bibliography

[Tandon et al.(2014)] Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In Proceedings of the 7th ACM international conference on Web Search and Data Mining (WSDM 2014).

[Sap et al.(2019)] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI conference on artificial intelligence (AAAI 2019).

[Romero et al.(2019)] Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019).

[Nguyen et al.(2021)] Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In Proceedings of the Web Conference 2021 (WWW 2021).

[Elazar et al.(2019)] Yanai Elazar et al. 2019. How Large Are Lions? Inducing Distributions over Quantitative Attributes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).

[Darari et al.(2015)] Fariz Darari, Radityo Eko Prasojo, and Werner Nutt. 2015. Expressing novalue information in RDF. In Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC 2015).

[Hoffart et al.(2013)] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial intelligence (2013).

[Singhania et al.(2023)] Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. Extracting Multi-valued Relations from Language Models. In Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023).

Bibliography

[Radford et al. (2018)] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[Devlin et al. (2019)] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT.

[Raffel et al. (2020)] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In Journal of Machine Learning Research.

[Du et al. (2022)] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACL 2022).

[Petroni et al. (2019)] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019).

[Mirza et al.(2018)] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2018. Enriching Knowledge Bases with Counting Quantifiers. In International Semantic Web Conference (ISWC 2018).

[Bast and Haussmann(2015)] Hannah Bast and Elmar Haussmann. 2015. More accurate question answering on Freebase. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM 2015).

Bibliography

[Izacard and Grave(2021)] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL 2021).

[Yu et al.(2022)] Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2022).

[Rajpurkar et al.(2018)] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2018).

[Rajpurkar et al.(2016)] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (EMNLP 2016).

[Dubey et al.(2019)] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A large dataset for complex question answering over Wikidata and DBpedia.

[Fan et al.(2019)] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).

[Lewis et al.(2020)] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems (2020).

Bibliography

[Day(2015)] Charles Day. 2015. One million physicists. (2015).

[NobelPrize(2023)] All Nobel Prizes in Physics. https://www.nobelprize.org/prizes/lists/all-nobel-prizes-in-physics. Last retrieved in 2023.

[Galárraga et al.(2017)] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M Suchanek. 2017. Predicting Completeness in Knowledge Bases. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017).

[Mirza et al.(2017)] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2017. Cardinal Virtues: Extracting Relation Cardinalities from Text. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2017).

[Soulet et al.(2018)] Arnaud Soulet, Arnaud Giacometti, Béatrice Markhoff, and Fabian M Suchanek. 2018. Representativeness of Knowledge Bases with the Generalized Benford's Law. In International Semantic Web Conference (ISWC 2018).

[Razniewski et al.(2019)] Simon Razniewski, Nitisha Jain, Paramita Mirza, and Gerhard Weikum. 2019. Coverage of Information Extraction from Sentences and Paragraphs. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019).

[Luggen et al.(2019)] Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. Non-parametric class completeness estimators for collaborative knowledge graphs—the case of wikidata. In International Semantic Web Conference (ISWC 2019).

[Baidu(2020)] Baidu. 2020. Introducing Qian Yan, Baidu's New Plan to Build 100 Chinese NLP Datasets in Three Years. http://research.baidu.com/Blog/index-view?id=146.

Bibliography

[Diefenbach et al.(2018)] Dennis Diefenbach, Vanessa L´opez, Kamal Deep Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. (2018).

[Karpukhin et al.(2020)] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).

[Qian(2013)] Richard Qian. 2013. Understand Your World with Bing. https://blogs.bing.com/search/March-2013/Understand-Your-World-with-Bing.