Class Cardinality Comparison as a Fermi Problem

Shrestha Ghosh, Simon Razniewski, Gerhard Weikum
Max Planck Institute for Informatics, Germany

Motivation

- **Incompleteness and bias in online world**
  Different sources focus on different angles
- **Cardinality signals have partial coverage**
  SE return American lawyers and board game types
- **Smart human's approach towards reasonable estimates**
  Inspired by Enrico Fermi [1]

Are there more **lawyers** than **police officers**?
more **board games** than **satellites**?

Sources		Sources	
Wikidata	96K > 8K	Wikidata	2.9K < 6.7K
SE results	1.3M > 18K	SE results	17 < 2.6K
GPT-3	1.3M = 1.3M	GPT-3	1.5K < 6K
Ground-truth	3.5M < 13M	Ground-truth	86K > 4.8K

Images credits: https://openclipart.org/

Problem Statement: Is class A bigger than class B

Cardinality Estimation *compute |A| and |B|* → Proportionality Estimation *compute |A|:|B|* → Dominance Estimation *compute |A|>|B|*

More reliable, more certainty

$$\text{Output}_{|A|>|B|} = \begin{cases} 1, & \text{if predict } |A| > |B| \\ -1, & \text{if predict } |A| < |B| \\ 0, & \text{if predictor abstains} \end{cases}$$

Cardinality Signal Extraction

by Sources

- **Knowledge Base** *Wikidata SPARQL queries*
- **Search Engine** *Inference over top-50 SE results using CoQEx [2]*
- **Language Model** *Extract class cardinality from GPT-3 answers*

by Types

- **Root**: *Count of class itself*
- **Subgroup**: *Counts of subgroups e.g. by country*

Cardinality Signal Aggregation

1. Subgroup Aggregation

> θ_M % subgroups of A have more entities than the those of B. — Majority

Subgroup cardinalities of A is significantly greater than that of B — T-test

2. Type Ensembles

Ensemble^{source}: Root, Majority, T-test

3. Source Ensembles

Majority Vote, Weighted Vote: Ensemble^{KB}, Ensemble^{SE}, Ensemble^{LM}

Our Curated Dataset: 90 classes; 6 domains

#Class pairs = 4005
#In-domain pairs = 630
#Cross domain pairs = 3375

Lower cardinality ratio → More difficult prediction

Domain	Examples
creative work	film, board game, book
geographical entities	lake, castle, dam
man-made object	satellite, submarine
occupation	politician, actor, physicist
organization	university, football club
species	snake, insect, fish

Dataset & Results

Results

Source	Root (1)	Type Ensembles		
		(1)+(2)	(1)+(3)	(1)+(2)+(3)
KB	64.7	61.5	**65.9**	61.8
SE	65.4	65.7	65.4	**68.4**
LM	74.4	77.1	75.8	**79.4**
Source (KB, SE, LM) Ensembles				
Majority Vote	77.8	76.8	78.7	**78.9**
Weighted Vote	78.2	79.3	**83.7**	81.3
Non-expert human baseline				
Closed-book	75.0			
Open-book	76.0			

Accuracy of Root Signals and Ensembles
(1) Root; (2) Majority; (3) T-test; Baselines

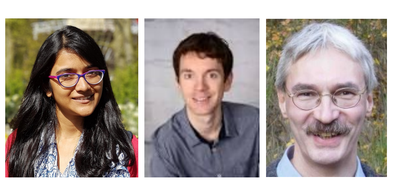
Type and source ensembles increase accuracy by 9.3% over root signals

Accuracy by sources vary widely across domains

Domain	KB	SE	LM	Best
Creative work	62.8	54.2	**83.8**	LM
Geographical entity	**77.1**	60.9	70.4	KB
Man-made object	26.6	77.1	**96.1**	LM
Occupation	57.1	**80.0**	74.2	SE
Organization	58.0	72.3	**88.5**	LM
Species	61.9	**78.0**	63.8	SE
Interdomain	62.6	68.0	**79.4**	LM
All	61.8	68.4	**79.4**	LM

Accuracy of Type Ensemble by Domain and Source

[1] Fermi Problem. https://en.wikipedia.org/wiki/Fermi_problem.
[2] Ghosh et al. 2022. Answering Count Queries with Explanatory Evidence. In SIGIR
Our research page on Count Knowledge: https://t1p.de/count_knowledge_research
Dataset and Results: https://github.com/ghoshs/class_cardinality_comparison



Contact ghoshs@mpi-inf.mpg.de