
Beyond Aggregations: Understanding Count Information for Question Answering

— Shrestha Ghosh —



Max Planck Institute for Informatics

Saarland Informatics Campus



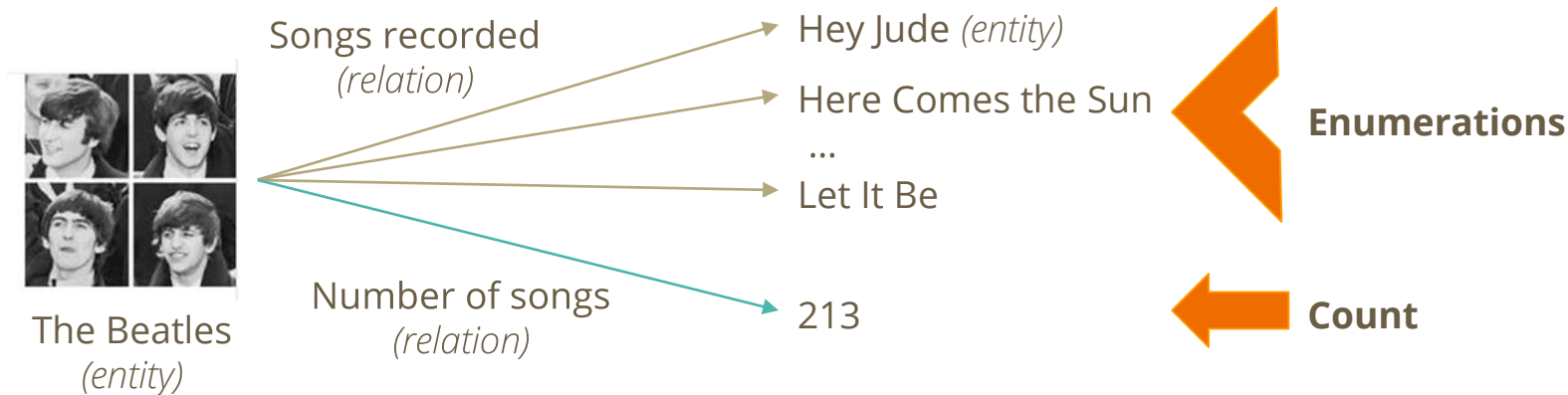
mpi
max planck institut
informatik



UNIVERSITÄT
DES
SAARLANDES



Count Information - What, Why and How?



Count Information - What, Why and How?

Popular QA datasets have **5%-10%** count related queries.

QA systems default to **ad-hoc count aggregation**.

121 languages (90% confidence)

Using data from 50 documents

Specific divisions:

1. 22 scheduled, 99 non-scheduled
2. 21 Indo-European, 17 Dravidian, ..

Sample enumerations:

Assamese, Bengali, Gujarati, Hindi, ..

A few months ago ..



A few weeks ago ..

22 languages

The Eighth Schedule of the Constitution consists of the following **22 languages** – Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu, Bodo, Santhali, Maithili and Dogri. Jul 1, 2018

indianexpress.com - India

More than 19,500 mother tongues spoken in India: Census ...

Goal

Number of languages in India



Related Work

1. Count information in IE
 - a. **Saha et al. (ACL 17)** - Bootstrapping for numerical open IE.
 - b. **Mirza et al. (ISWC 18)** - Enriching KB with counting quantifiers.
 - c. **Ho et al. (ISWC 19)** - Qsearch: Answering quantity queries from text.
2. Count information QA
 - a. **Abujabal et al. (ACL 17)** - QUINT: Interpretable QA over KBs.
 - b. **Bast et al. (CIKM 15)** - More accurate QA on Freebase.
 - c. **Diefenbach et al. (WWW 19)** - QAnswer (bridging the gap between a the LOD and end-user)
3. KB recall
 - a. **Paulheim (SWJ 17)** - KG refinement: A survey of approaches and evaluation methods



Research Question

Given a natural language (count) query:

- Number of songs by the Beatles
- Number of languages in India

*Provide a **correct**, **informative** and **explainable** answer.*

- **Correct** - true count or a reasonable estimate
- **Informative** - representative enumerations.
- **Explainable** - relevant context used to derive the count.

(Methodology)

Q1. How to ***identify count information*** in text.

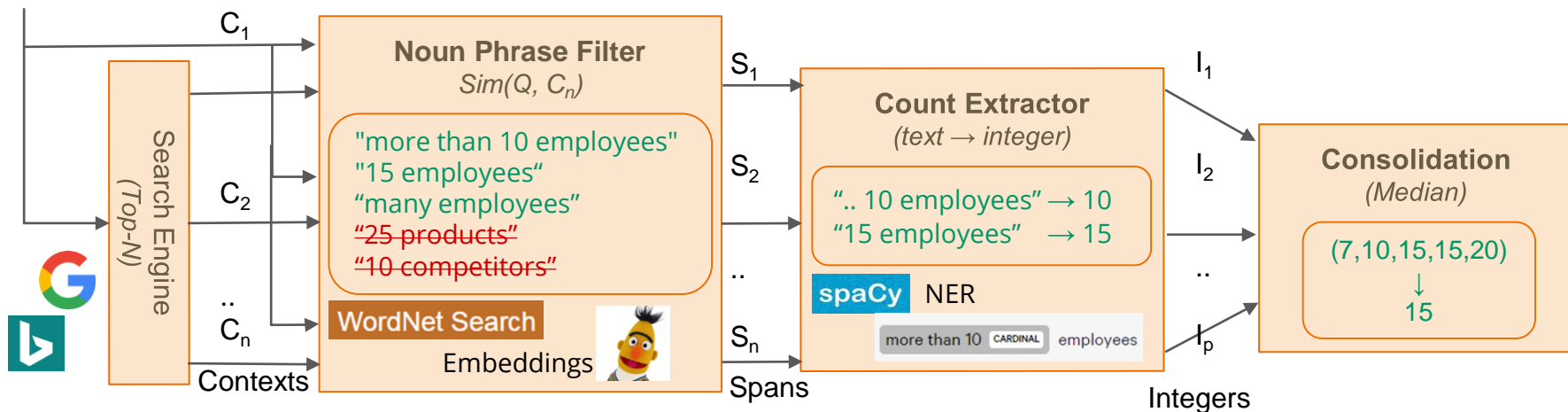
(Evaluation)

Q2. How to ensure a) correctness b) informativeness and c) explainability.

Methodology - Count Extraction

Query

(Q: number of employees in Microsoft)



Text Retrieval

Count Extraction

Consolidation



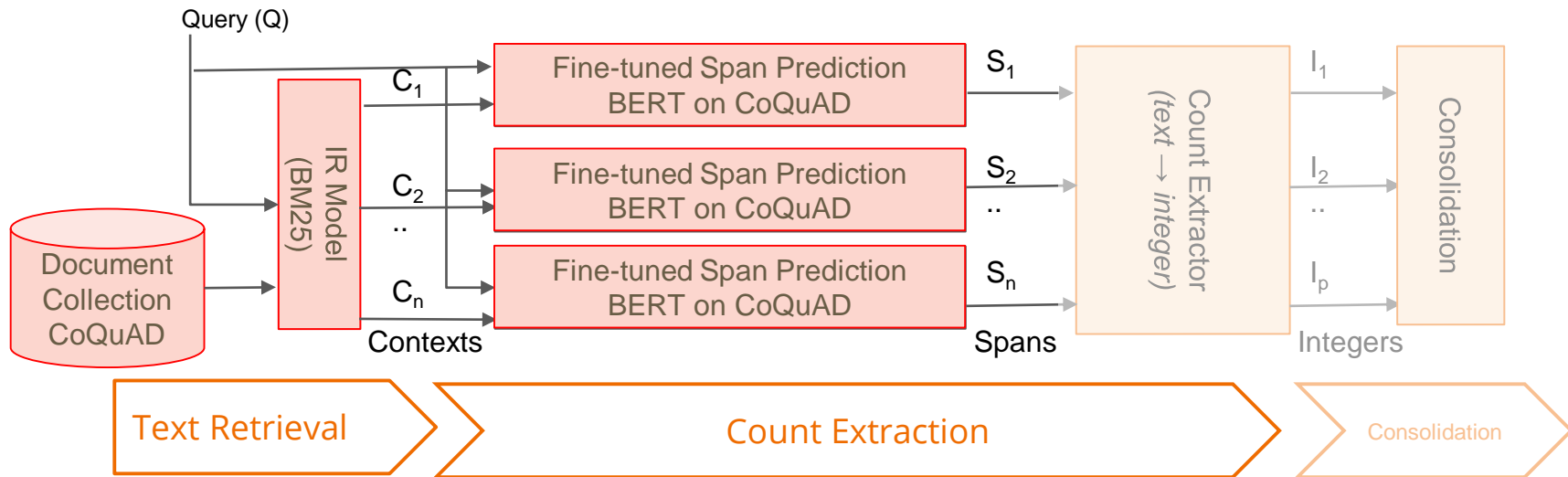
Evaluation and Preliminary Results





1. Precision - fraction of predictions which exactly match the gold answer.
2. Correctness ratio - ***min(prediction, gold)*** to the ***max(prediction, gold)***.
 - Measures proximity to the true count.
 - A tolerance threshold for a **relaxed precision** measure.
3. Normalized 95% confidence interval size - Compute ratio of the CI size to the predicted value.

Similarity (Q, C _n)	Precision	Rel. Precision	Correctness ratio	Norm. CI size
Head noun, Head noun	~0.3	~0.3	~0.5	(10 ³ - 10 ⁶)
Head noun, Noun phrase				
Query, Noun phrase				



Evaluation and Preliminary Results




Precision	Rel. Precision	Correctness ratio	Norm. CI size
~0.7 	>0.7 	>0.8 	<10 



Next Steps

1. Identify enumerations from text snippets
 - Named entities belonging to the same class as the queries noun
2. Evaluate the quality of enumerations

India GPE has a Greenberg's diversity index of 0.914—i.e. CARDINAL . two CARDINAL people selected at random from the country will have different native languages in 91.4% of cases. [8 CARDINAL] As per the 2011 Census of India GPE , languages by highest number of speakers are as follows: Hindi , Bengali , Marathi , Telugu , Tamil , Gujarati GPE , Urdu GPE , Kannada GPE , Odia GPE , Malayalam GPE .



Languages identified as a geo-political entity

Unidentified languages

Current state-of-the-art NER systems