CACAO

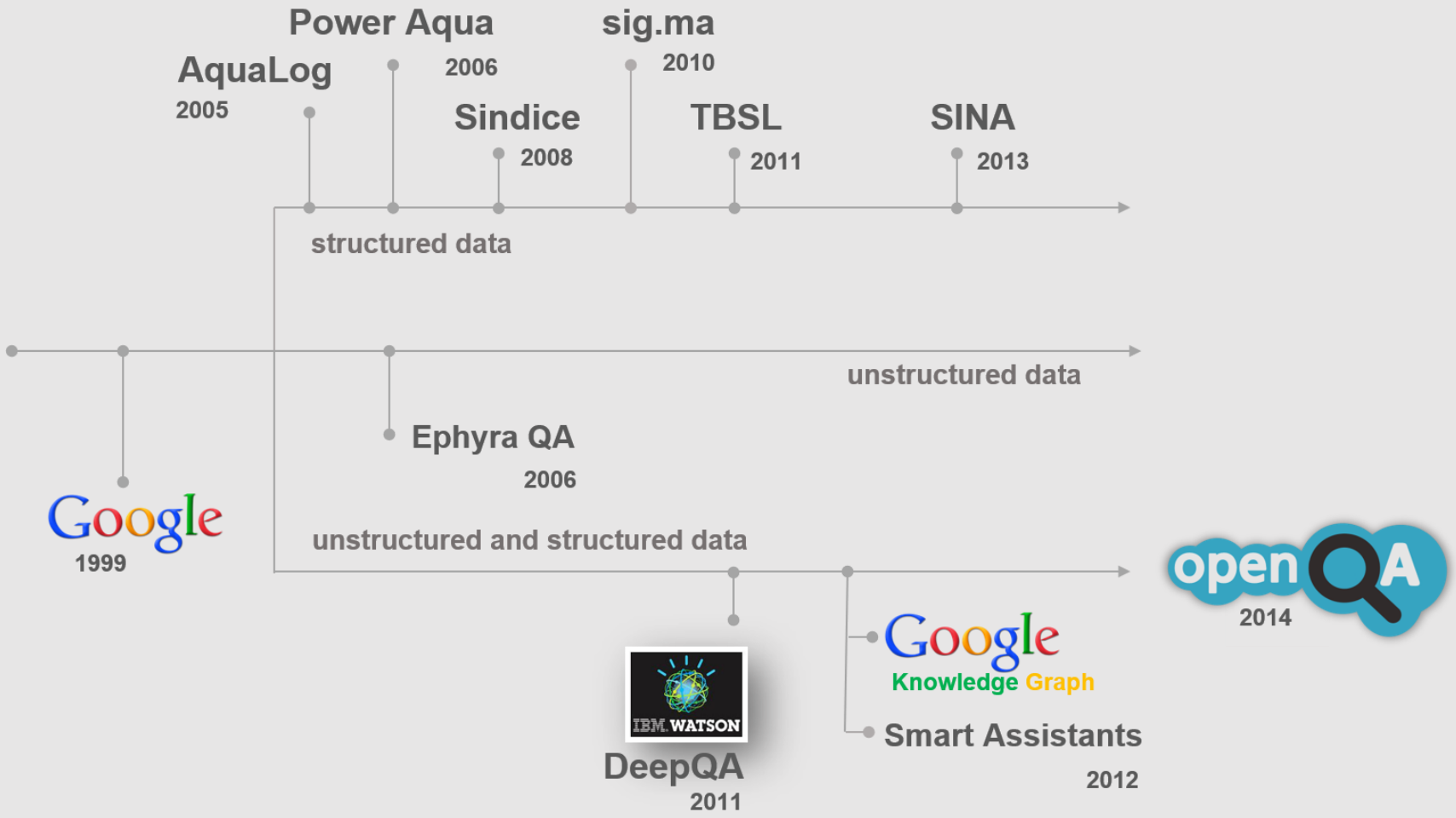
Conditional Spread Activation for Keyword Query Interpretation

15th Internation Conference on Semantic Web
Karlsruhe 2019

Edgard Marx, Gustavo Publio & Thomas Riechert

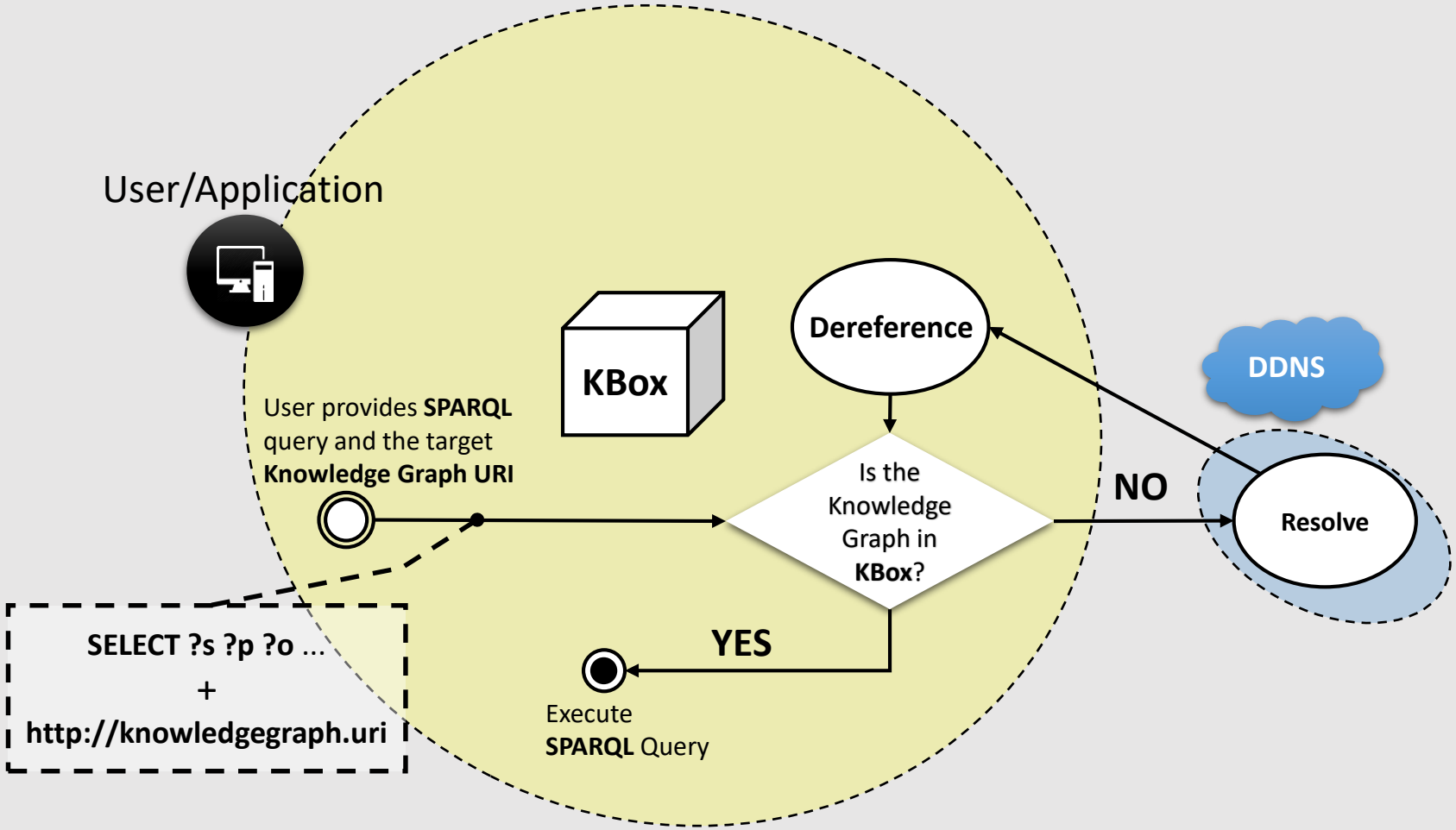
LiberAI HTWK Leipzig AKSW

openQA



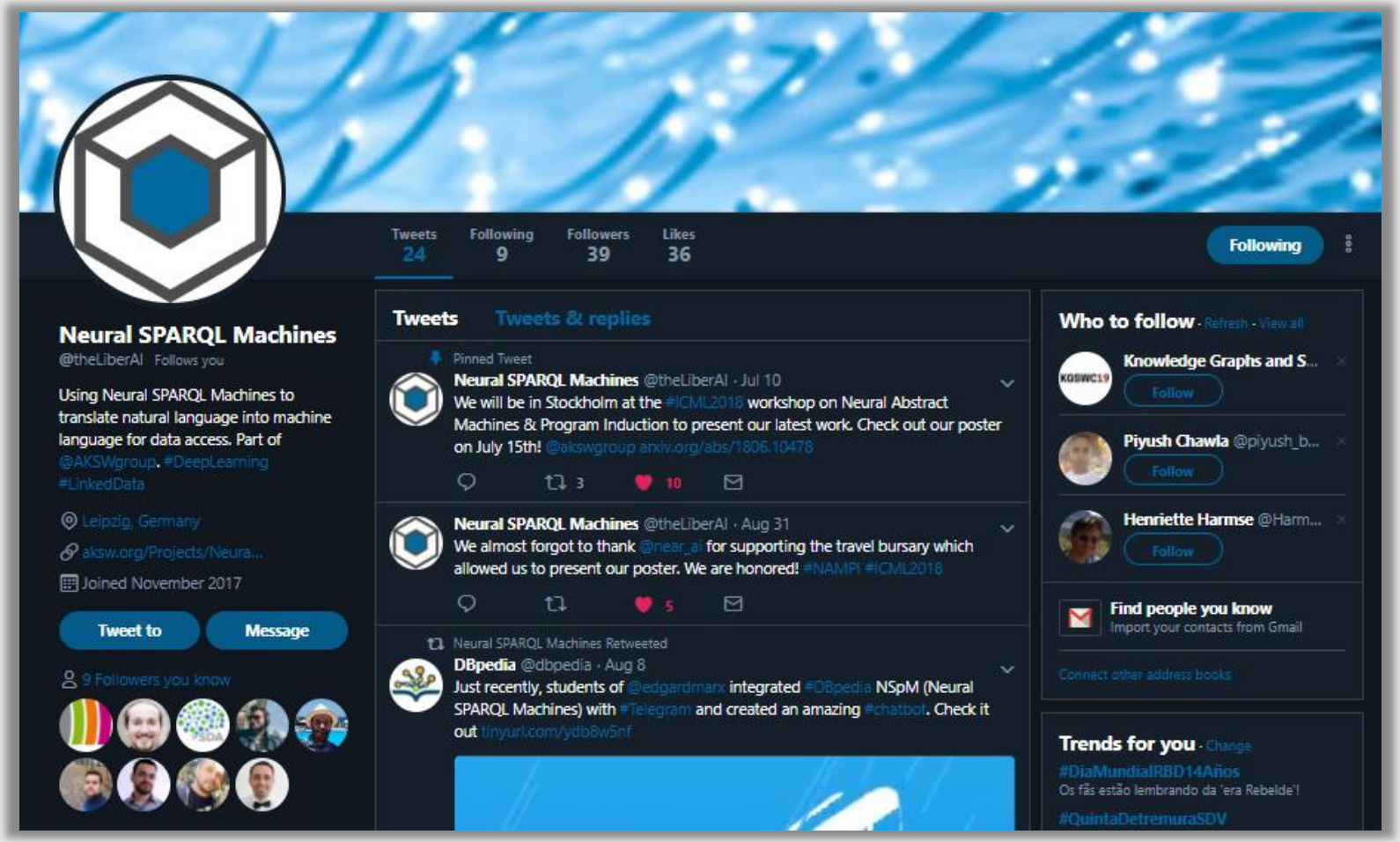
*https://github.com/AKSW/openQA

KBox



*https://github.com/AKSW/KBox

NSpM



*https://github.com/AKSW/NSpM

DBNQA

Question Answering data [edit]

This section includes datasets that deals with structured data.

Dataset Name	Brief description	Preprocessing	Instances	Format	Default Task	Created (updated)	Reference	Creator
DBpedia Neural Question Answering (DBNQA) Dataset	A large collection of Question to SPARQL specially design for Open Domain Neural Question Answering over DBpedia Knowledgebase.	This dataset contains a large collection of Open Neural SPARQL Templates and instances for training Neural SPARQL Machines; it was pre-processed by semi-automatic annotation tools as well as by three SPARQL experts.	894,499	Question-query pairs	Question Answering	2018	[402][403]	Hartmann, Soru, and Marx et al.

https://wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

*https://github.com/AKSW/DBNQA

Outline

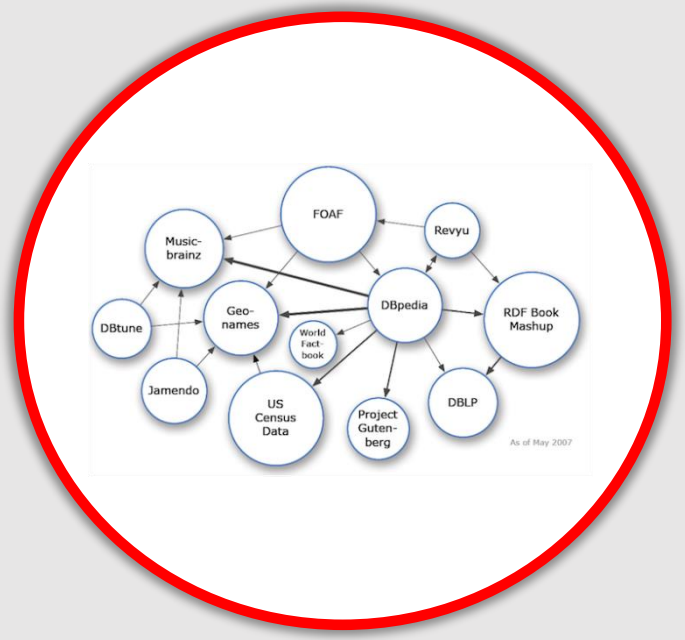
- Motivation
- Problem Statment
- Background
- Approach
- Evaluation
- Counclusion & Future Works

Linked Data

Unstructured

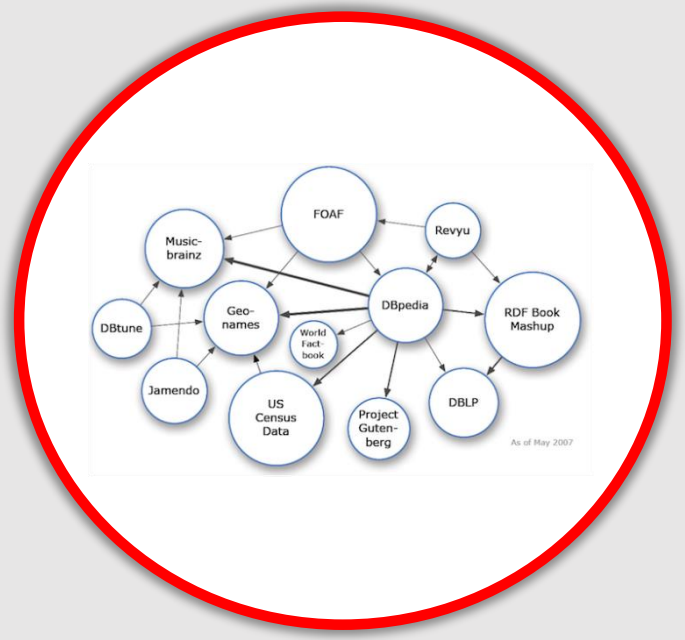
Structured

Crowd

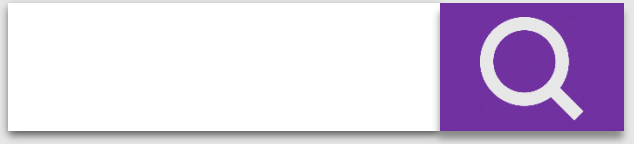


Linked Data

Access (the information needed)



Linked Data



- Entity Retrieval
- Question Answering
- Entity Linking

Entity Retrieval

Formally, a top-k entity retrieval takes a keyword query Q, an integer $0 < k$, a set of entities $E=\{e_1, e_2, \ldots, e_{|E|}\}$, and returns the top-k entities based on a scoring function S(Q, e)

Entity Retrieval

```
Select * where {
        ?s dbo:birthPlace dbpedia:Leipzig
}
```

TF-IDF

$$w_{t,d} = tf_{t,d} \cdot log \frac{|D|}{|\{d' \in D | t \in d'\}|}$$

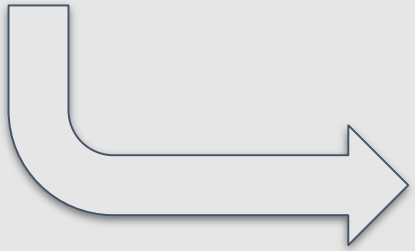
1: Winter is coming.
2: Ours is the fury.
3: The choice is yours.

term	freq	documents
choice	1	3
coming	1	1
fury	1	2
is	3	1, 2, 3
ours	1	2
the	2	2, 3
winter	1	1
yours	1	3

Dictionary | Postings

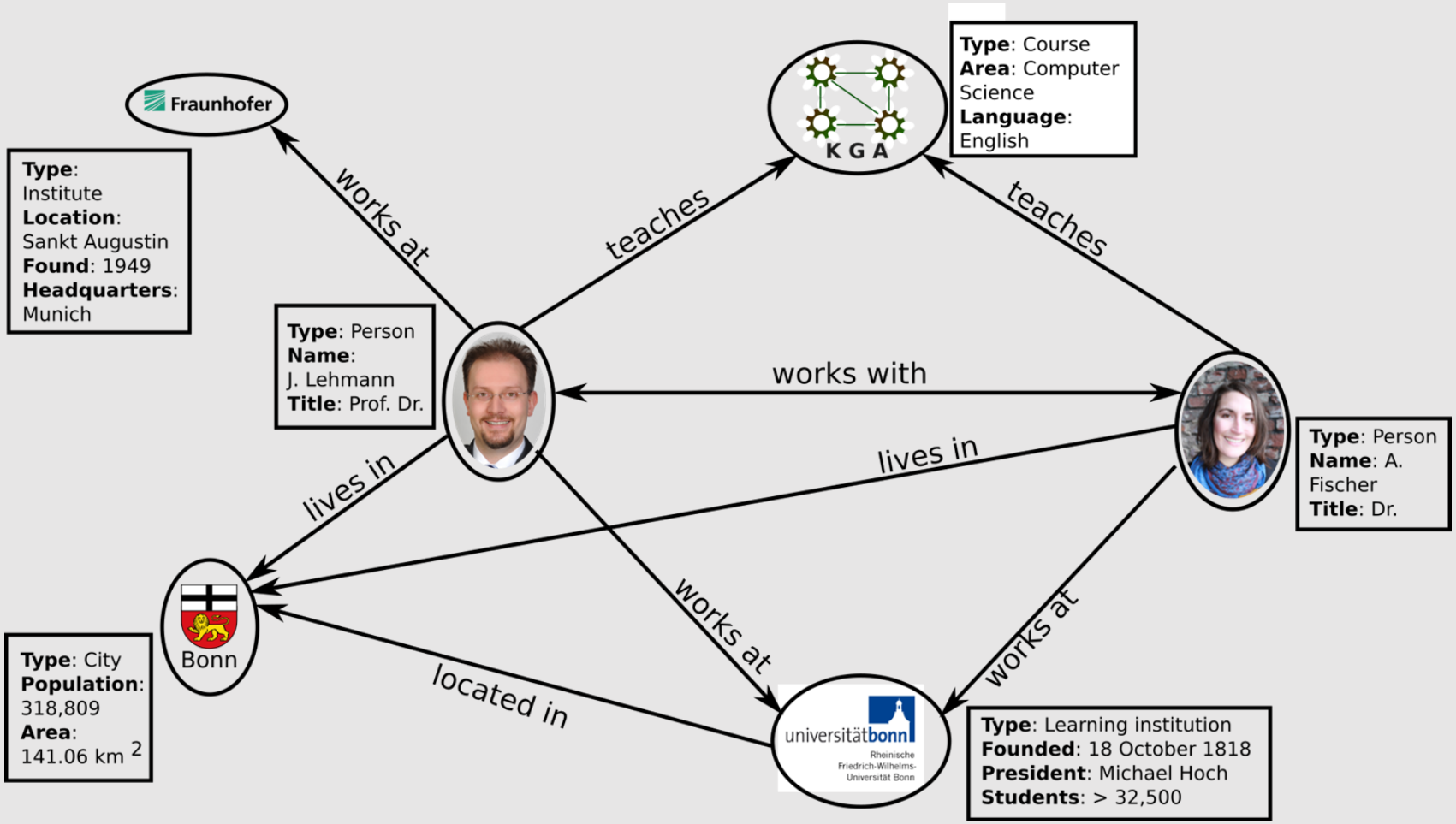
BM25

"This can opener can open any can that a can opener can and if this can opener cannot open any can that a can opener can, then this can opener is free"

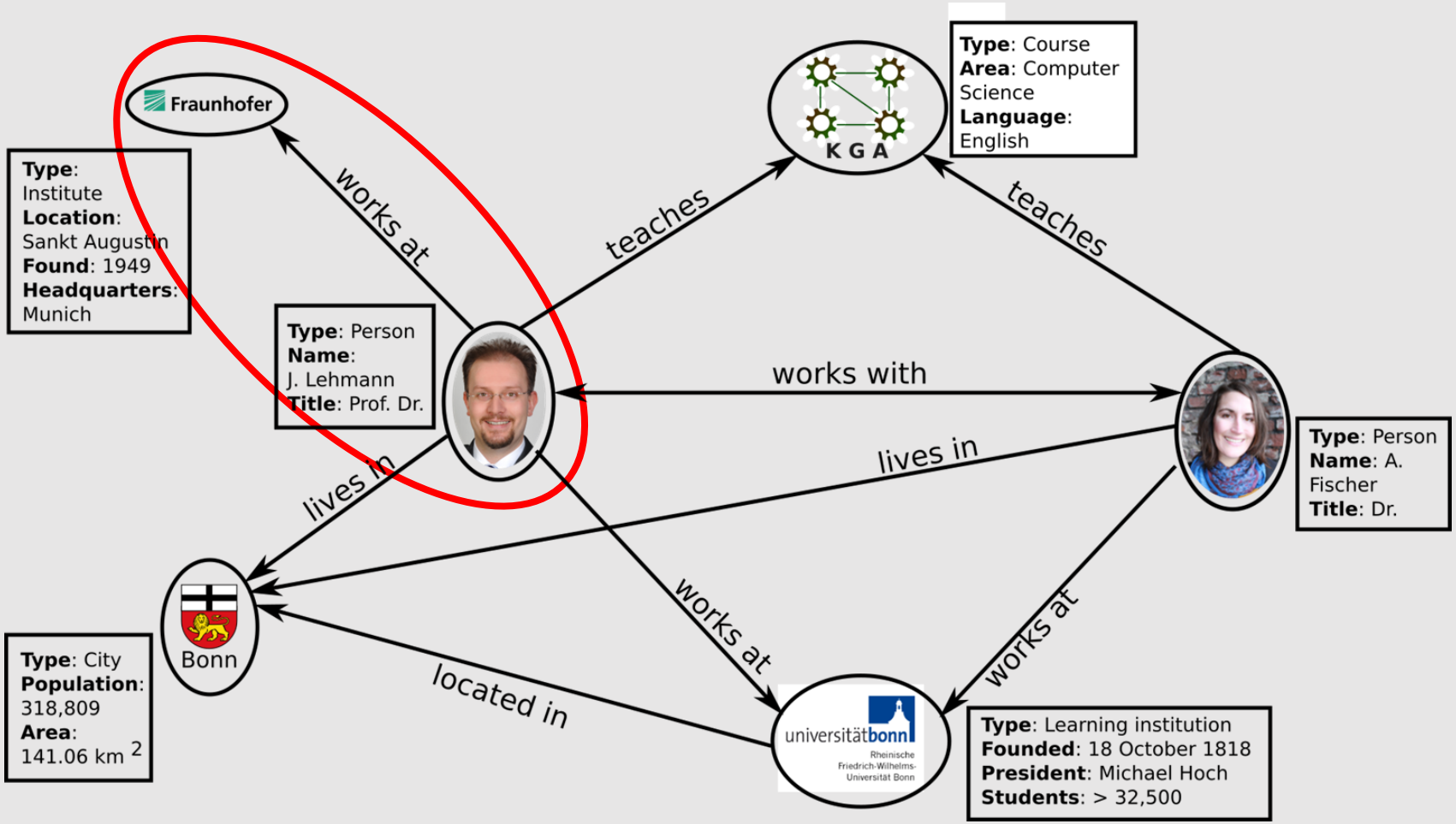


“If this can opener does not open any can, then you can take it for free”

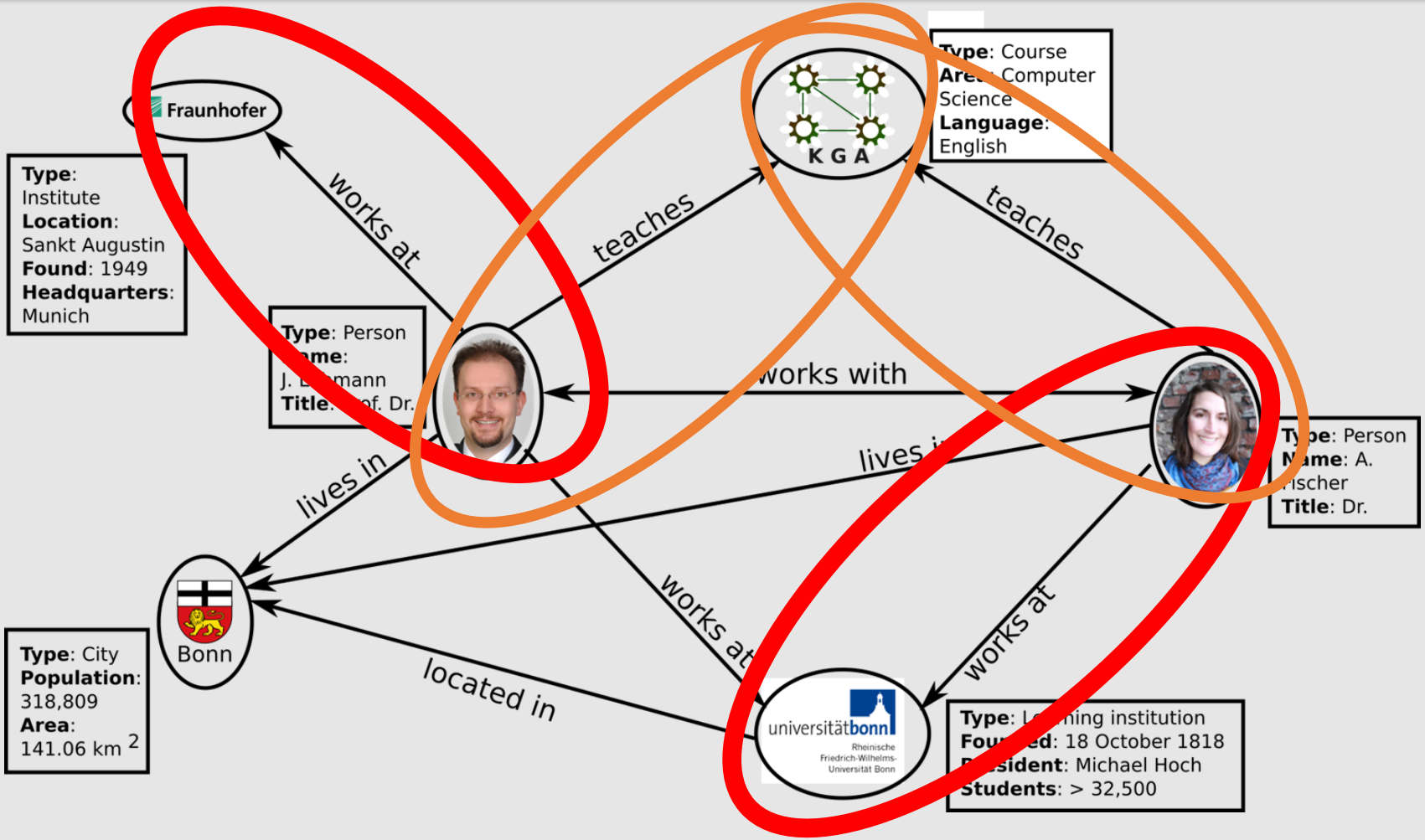
BM25


Fraunhofer
Type: Institute
Location: Sankt Augustin
Found: 1949
Headquarters: Munich
works at
teaches
K G A
Type: Course
Area: Computer Science
Language: English
teaches
Type: Person
Name: J. Lehmann
Title: Prof. Dr.
works with
lives in
lives in
Type: Person
Name: A. Fischer
Title: Dr.
Bonn
Type: City
Population: 318,809
Area: 141.06 km 2
works at
works at
located in
universitätbonn
Rheinische Friedrich-Wilhelms-Universität Bonn
Type: Learning institution
Founded: 18 October 1818
President: Michael Hoch
Students: > 32,500

BM25F

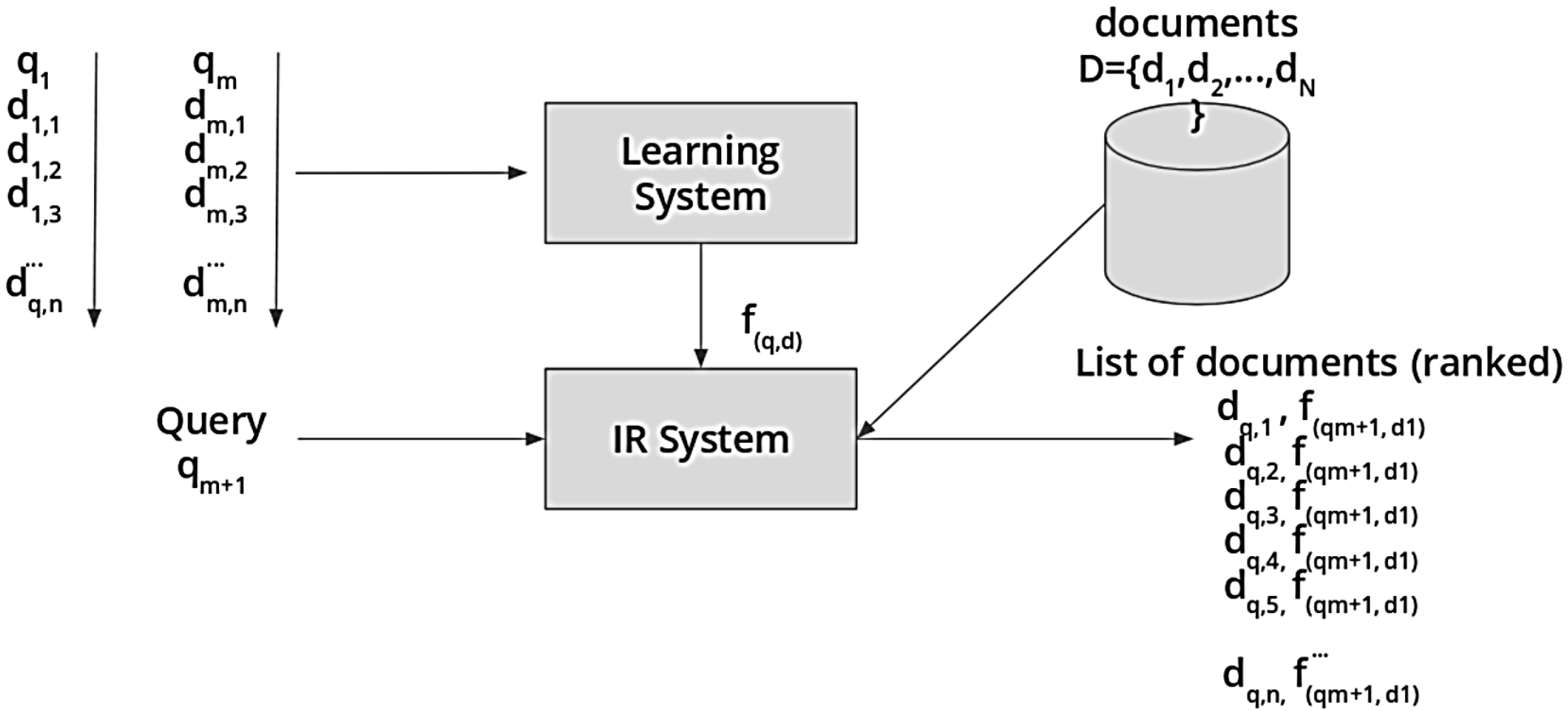

Fraunhofer
Type: Institute
Location: Sankt Augustin
Found: 1949
Headquarters: Munich
works at
Type: Person
Name: J. Lehmann
Title: Prof. Dr.
teaches
K G A
Type: Course
Area: Computer Science
Language: English
teaches
works with
lives in
lives in
Type: Person
Name: A. Fischer
Title: Dr.
Bonn
Type: City
Population: 318,809
Area: 141.06 km 2
located in
works at
works at
universitätbonn
Rheinische Friedrich-Wilhelms-Universität Bonn
Type: Learning institution
Founded: 18 October 1818
President: Michael Hoch
Students: > 32,500

Field Weighted

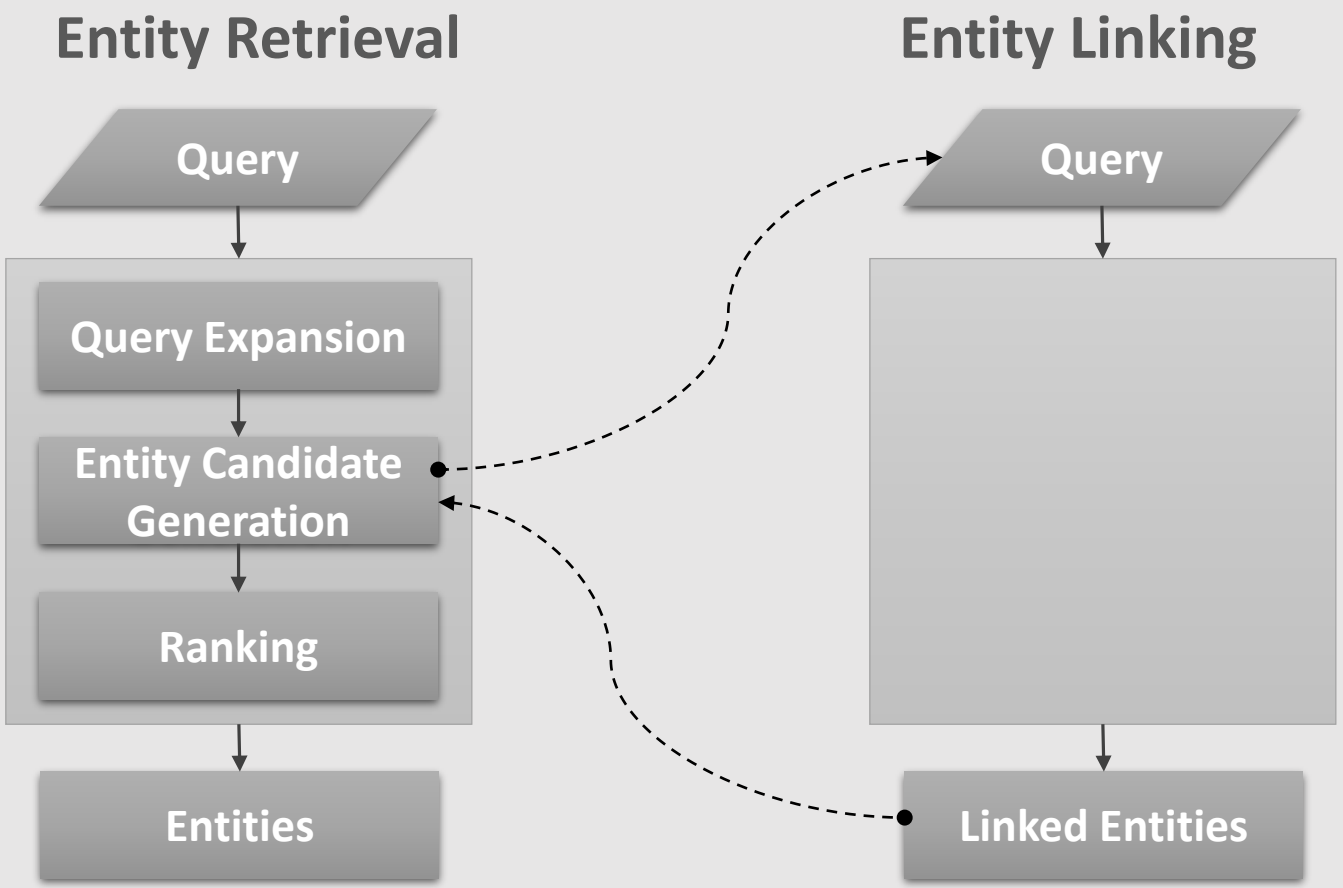


weight(p1) > weight(p2) > weight(p3) ...

Learning to Rank

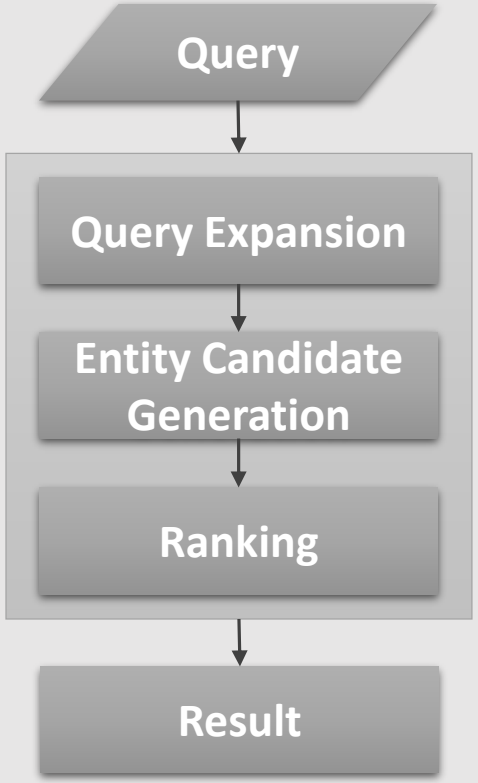


Entity Link in Queries

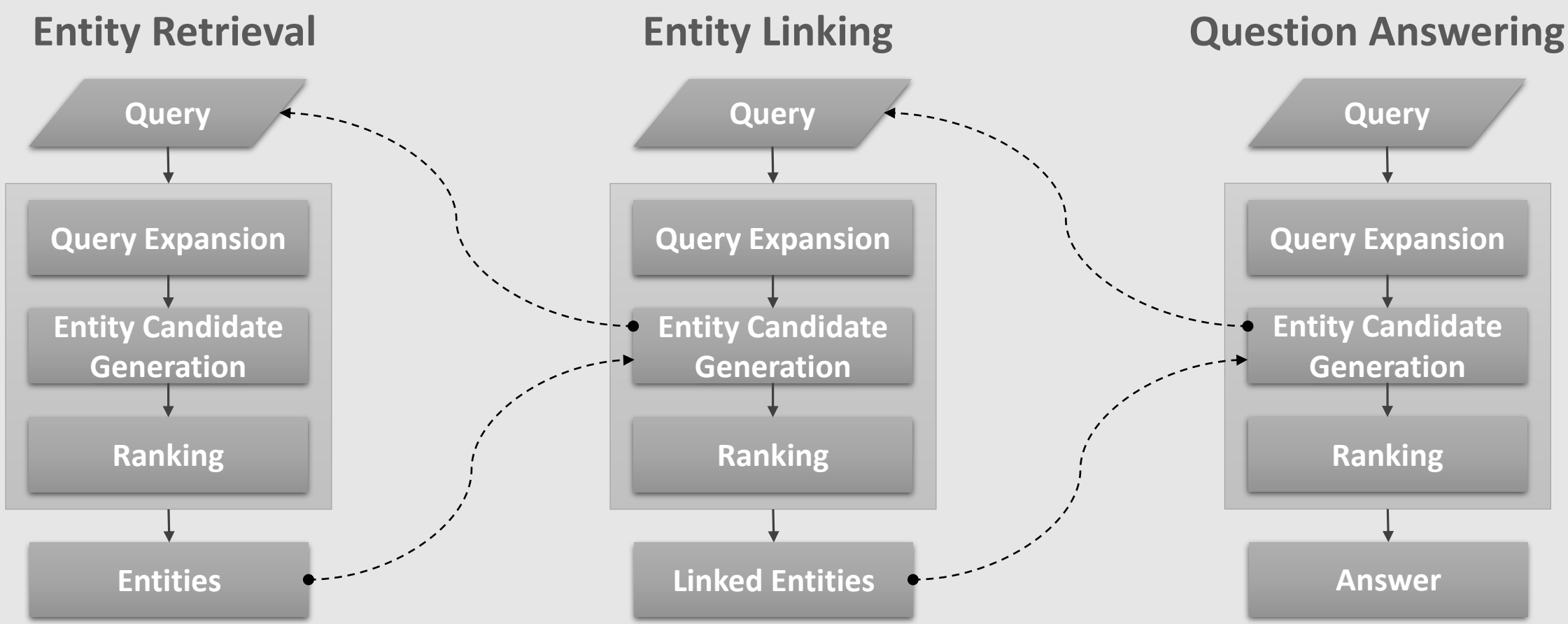


General IR Architecture

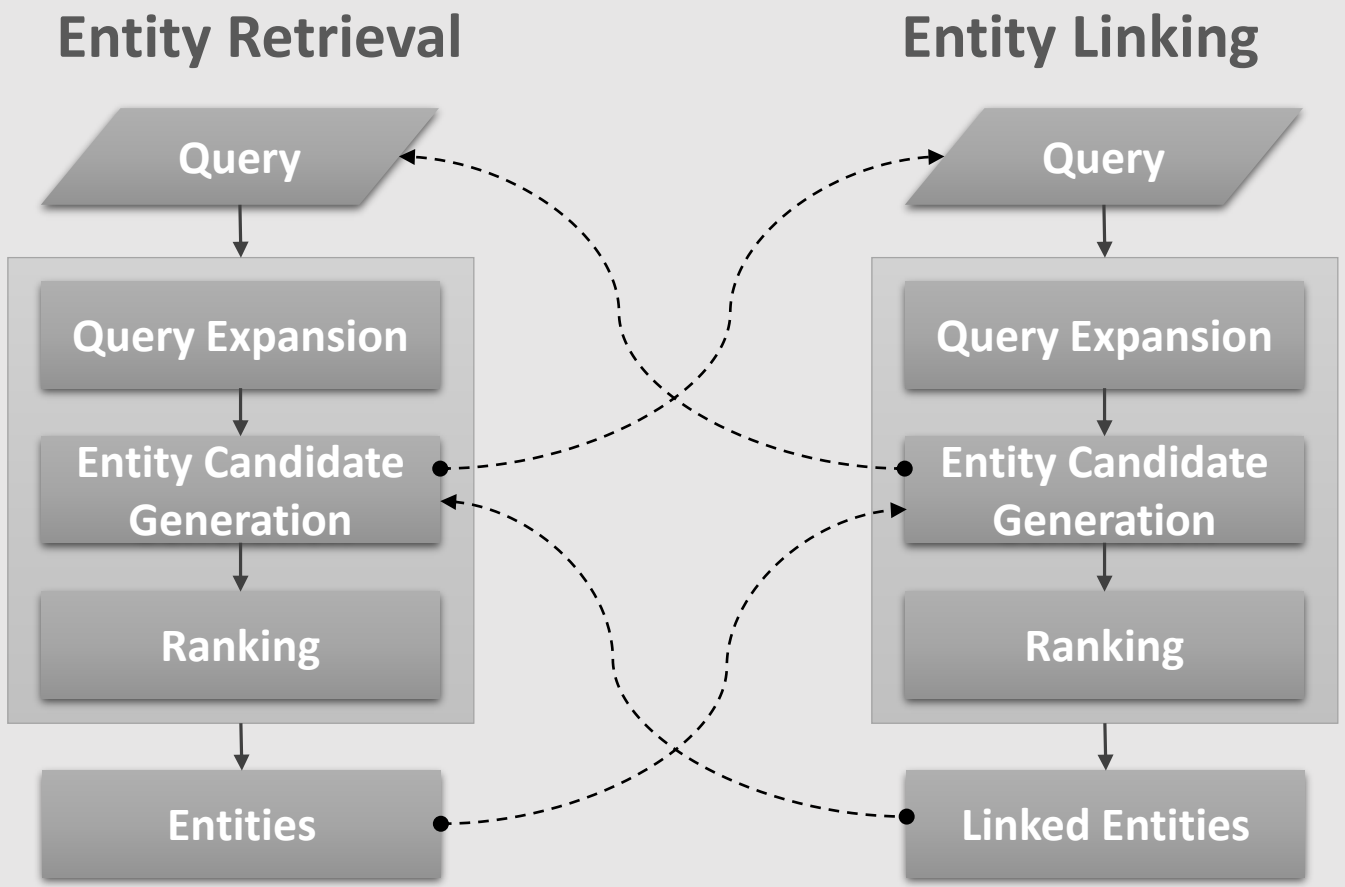
Entity Retrieval, Entity Linking, Question Answering

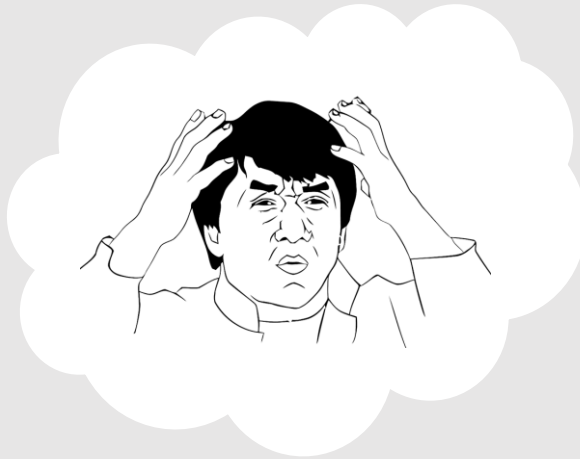


Interaction

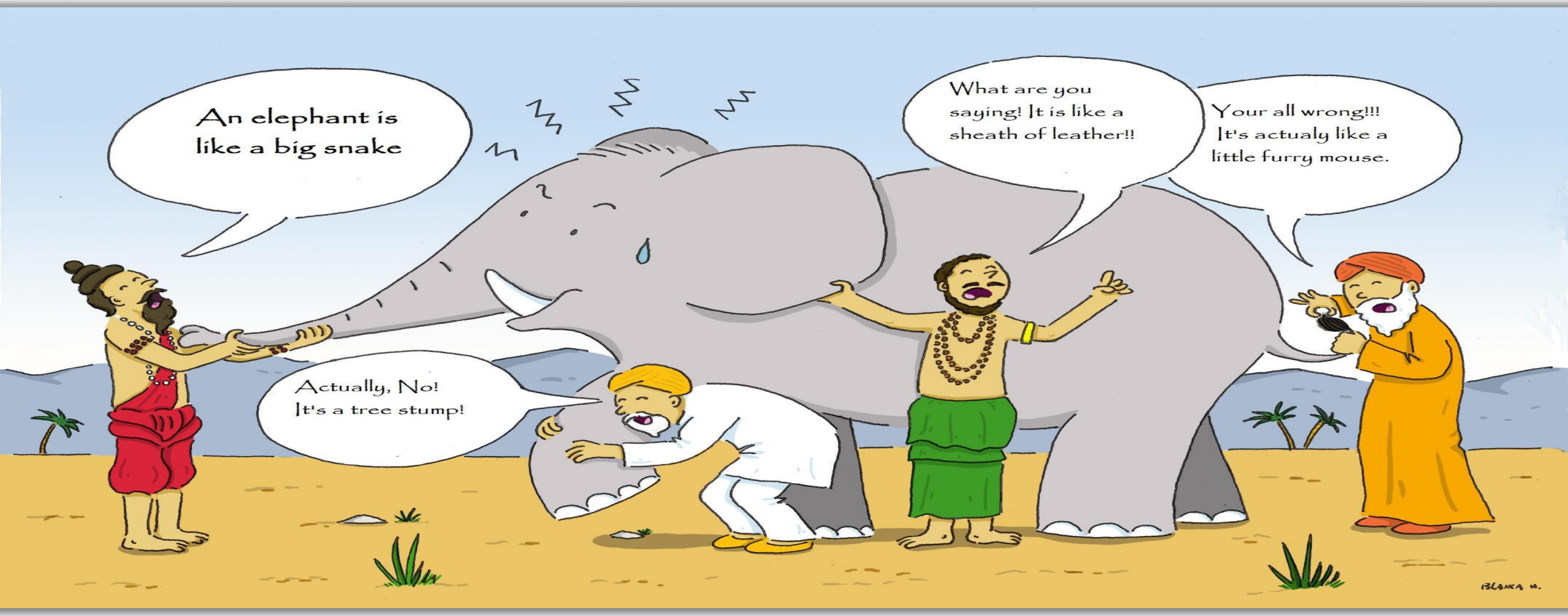


Lastest Architecture

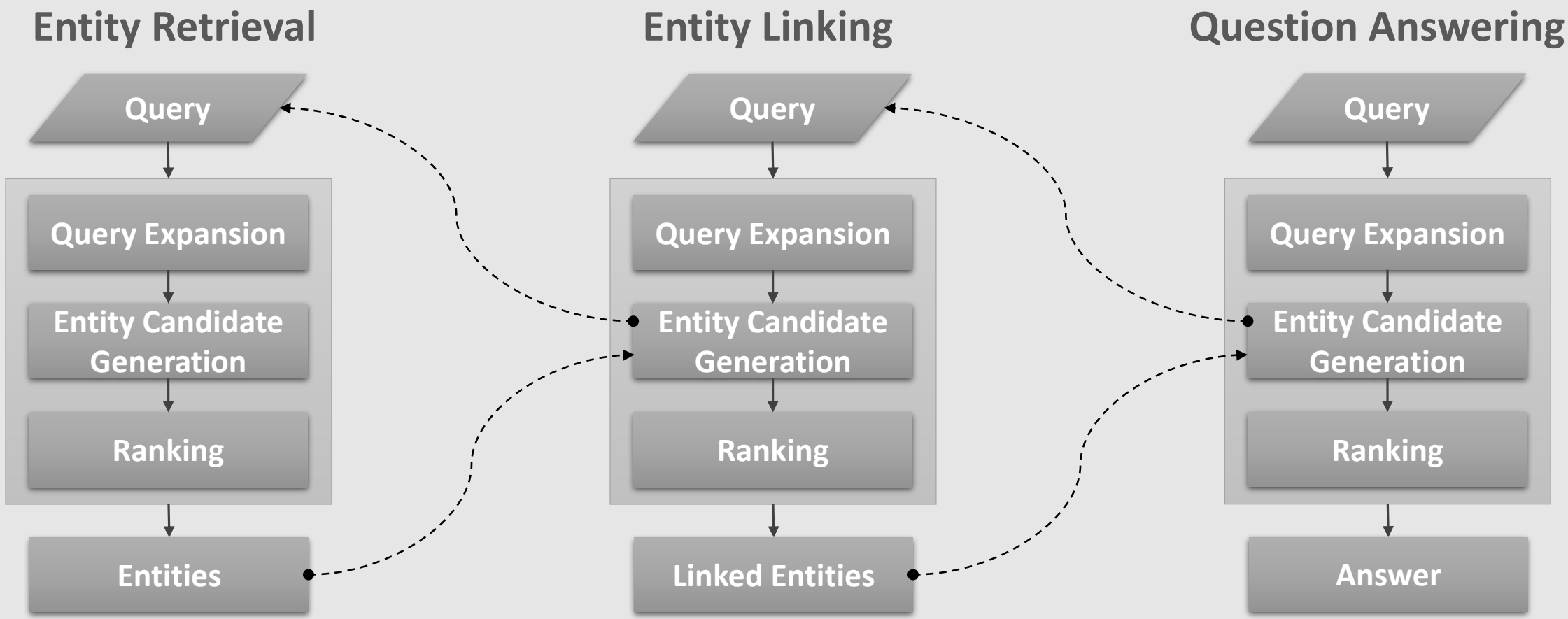




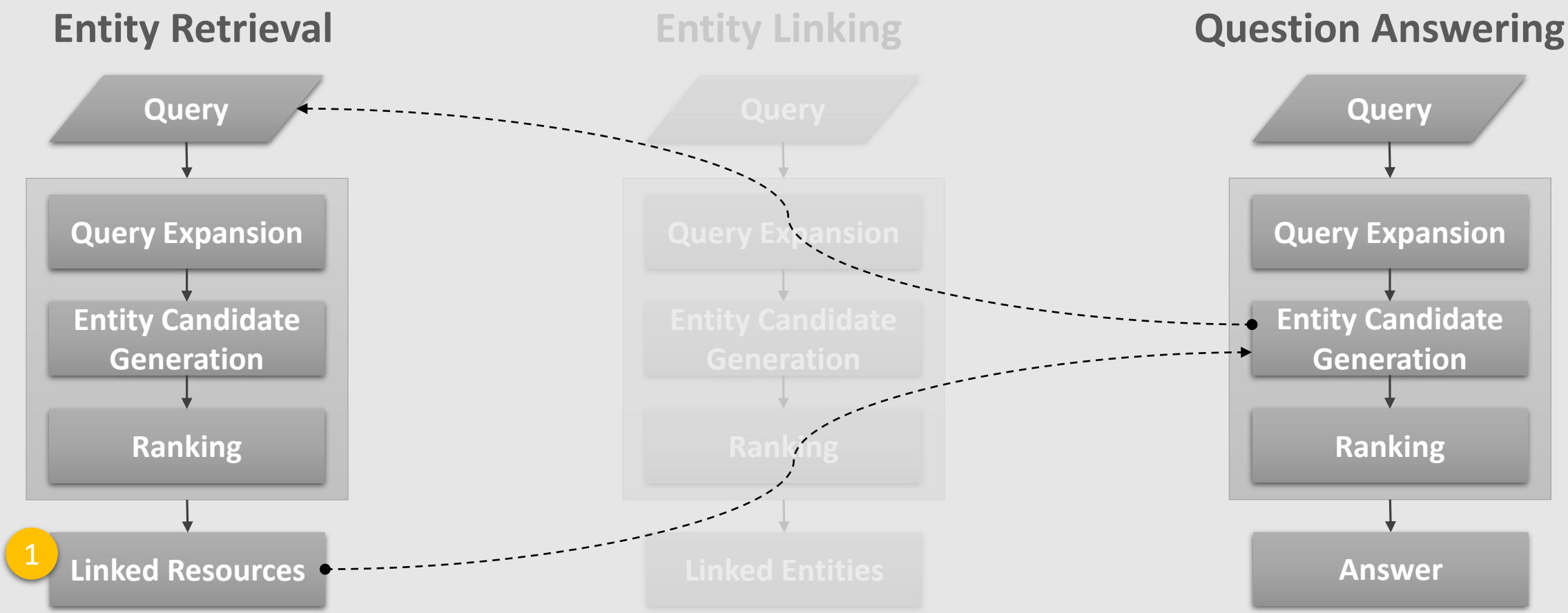
Lastest Architecture



Current Scenario

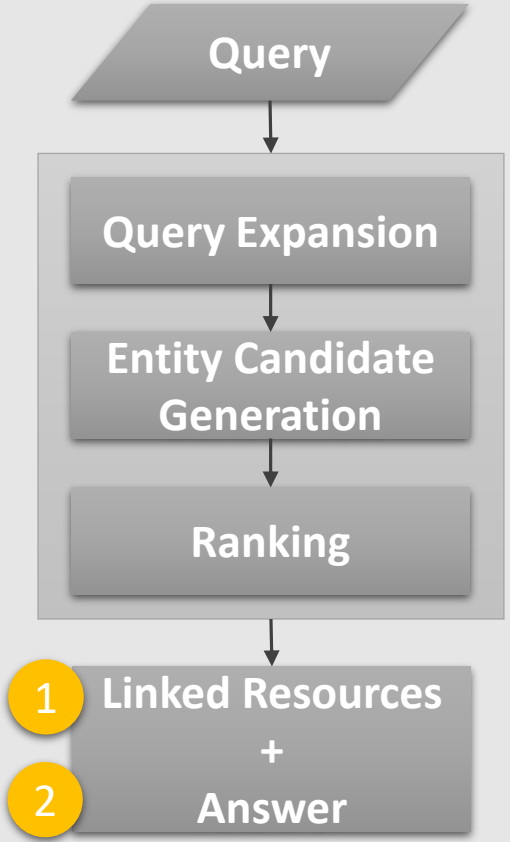


Return linked resources

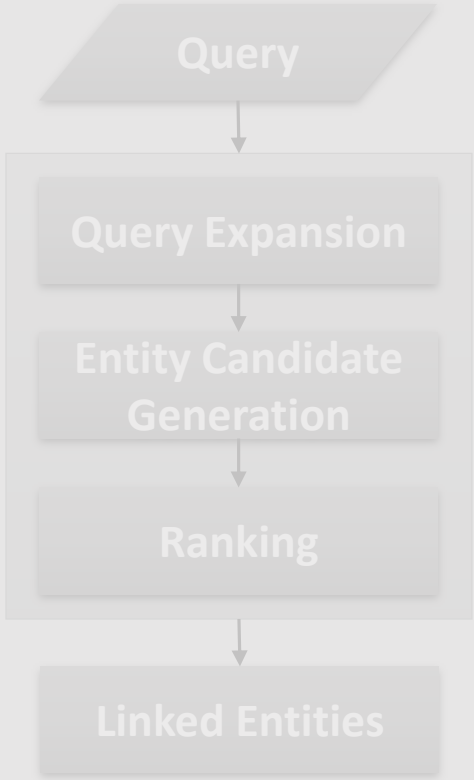


Return the answer

Entity Retrieval



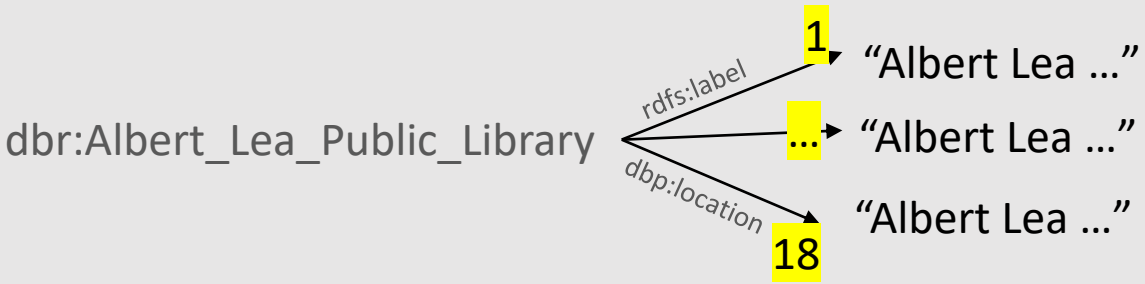
Entity Linking

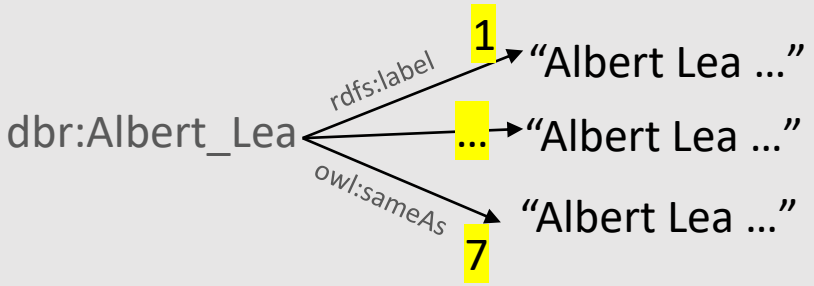


Question Answering

Challenge 1: *Local Term Frequency

Q: “Albert Lea”

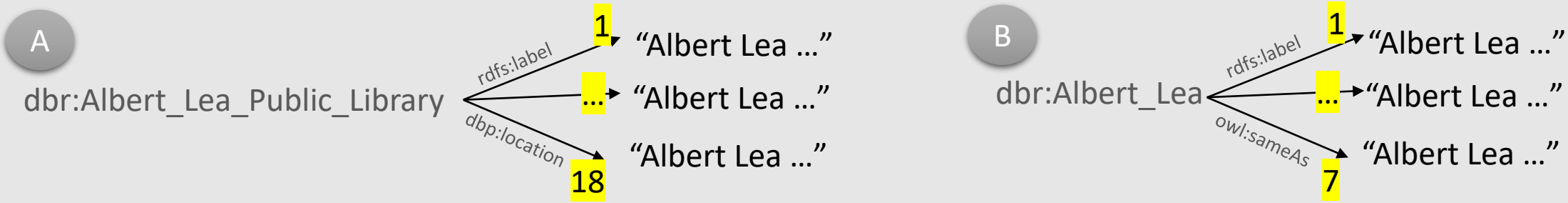




*not to confuse with global

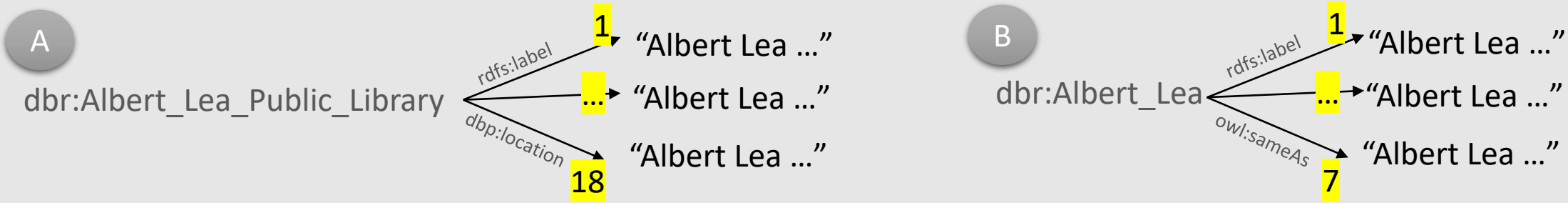
Challenge 2: Cohesion

Q: “Albert Lea”



Challenge 2: Cohesion

Q: “Albert Lea”

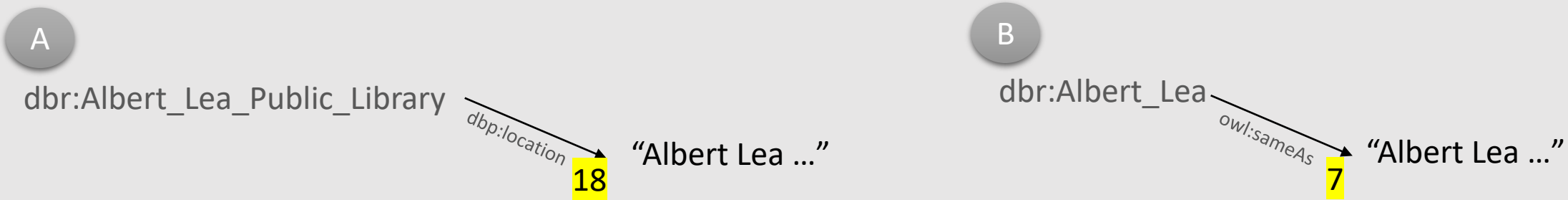


Preposition 1 [1][2]: maximize the number of tokens and reduce the number of mapped entities

[1] SWJ, Shekarpour et al. 2015, **[2]** NAACL, Sakor et al. 2019

Challenge 2: Cohesion

Q: “Albert Lea”



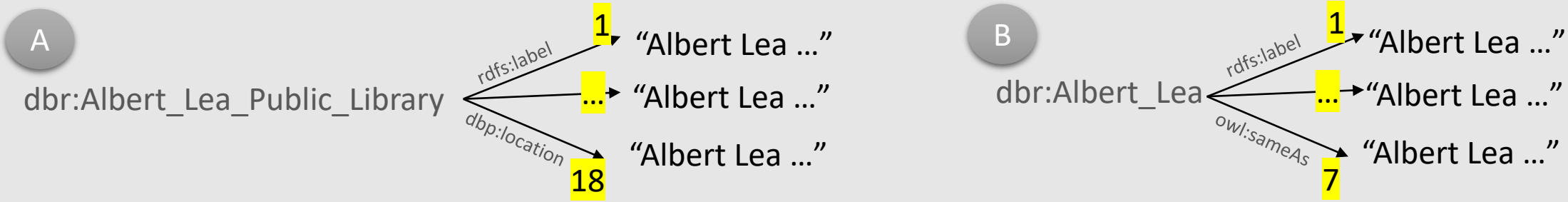
Proposition 1 [1][2]: maximize the number of tokens and reduce the number of mapped entities

Does **NOT** satisfy: Either A and B comply with the proposition 1

[1] SWJ, Shekarpour et al. 2015, **[2]** NAACL, Sakor et al. 2019

Challenge 2: Cohesion

Q: “Albert Lea”

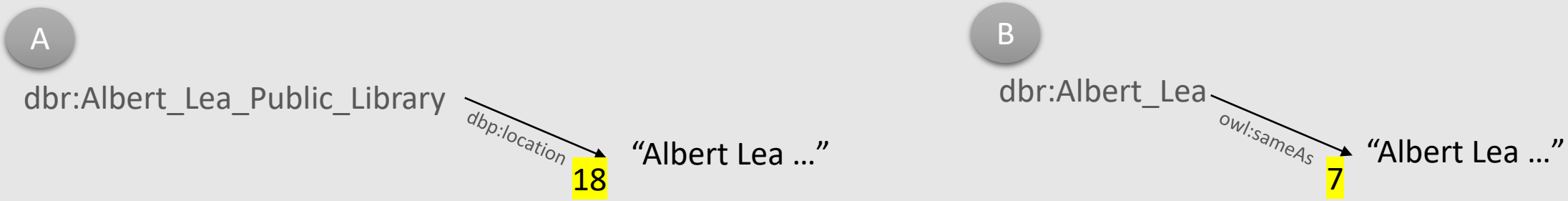


Proposition 2 [3][4]: use the context

[3] ISWC, Usbeck et al. 2014, **[4]** KCap, Moussallem et al. 2017

Challenge 2: Cohesion

Q: “Albert Lea”



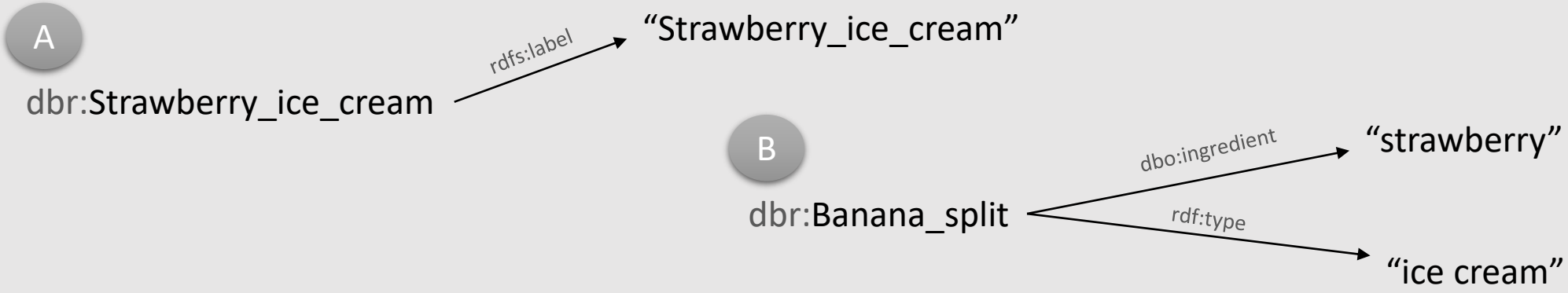
Proposition 2 [3][4]: use the context

Does **NOT** satisfy: Either A and B comply with preposition 2

[3] ISWC, Usbeck et al. 2014, **[4]** KCap, Moussallem et al. 2017

Challenge 2: Cohesion

Q: “strawberry ice cream”

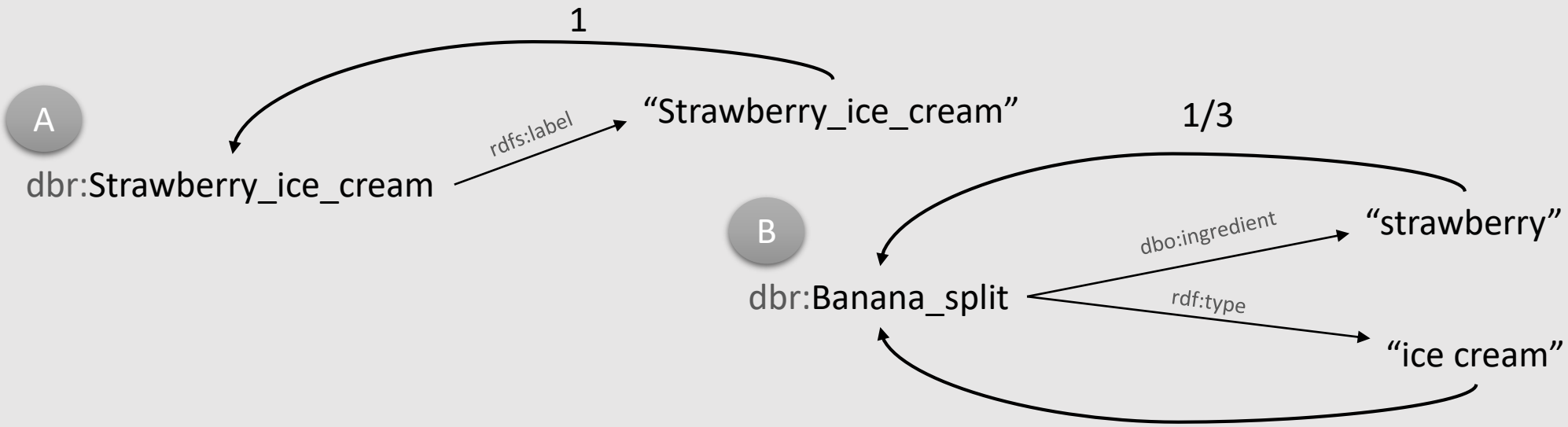


Proposition 3 [5][6]: Max Similarity (Levenshtein and Jaccard)

[5] AAAI, Zhang et al. 2016, **[6]** VLDB, Zheng et al. 2019

Challenge 2: Cohesion

Q: “strawberry ice cream”

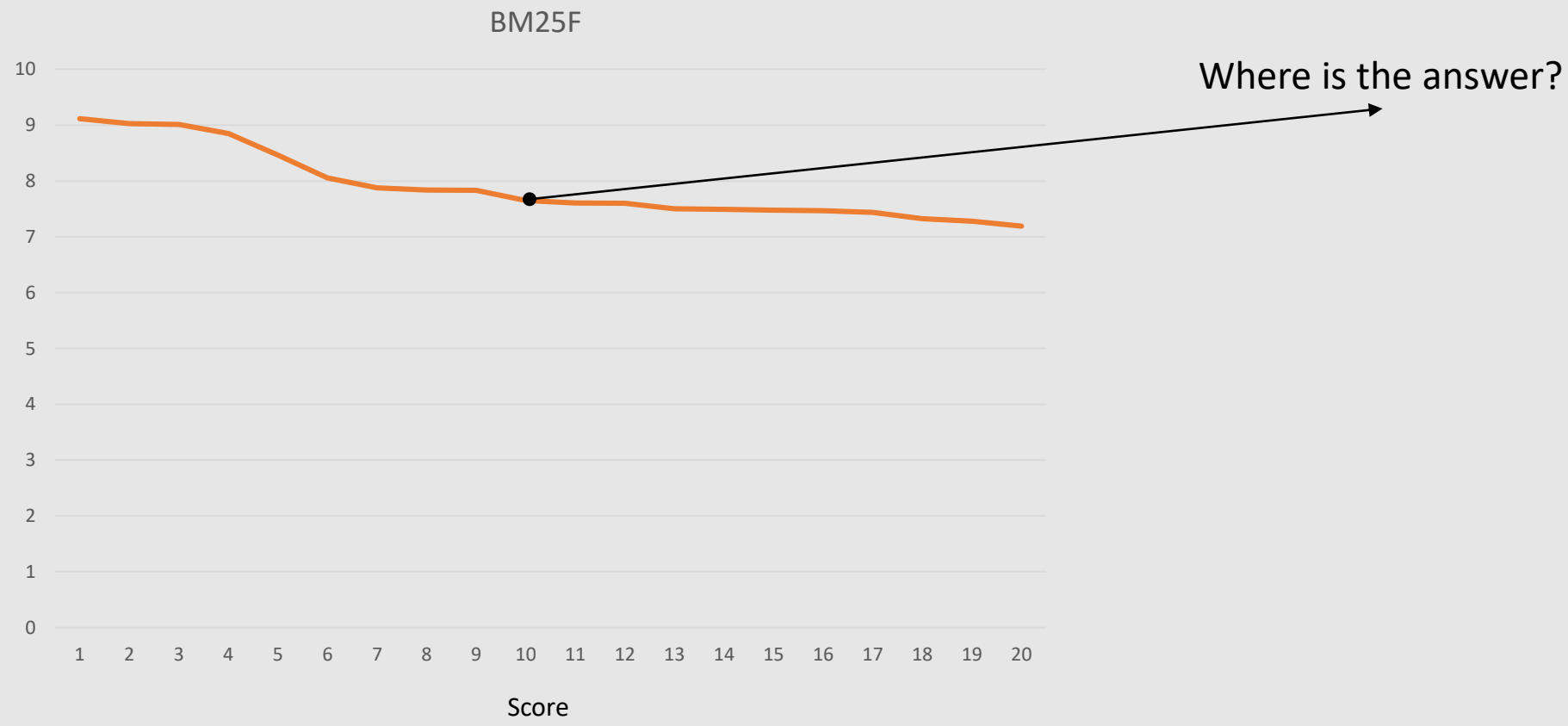


Proposition 3 [5][6]: Max Similarity (Levenshtein and Jaccard)

Does **NOT** satisfy: Either A and B comply with preposition 3

[5] AAAI, Zhang et al. 2016, **[6]** VLDB, Zheng et al. 2019

Challenge 3: Scoring Function



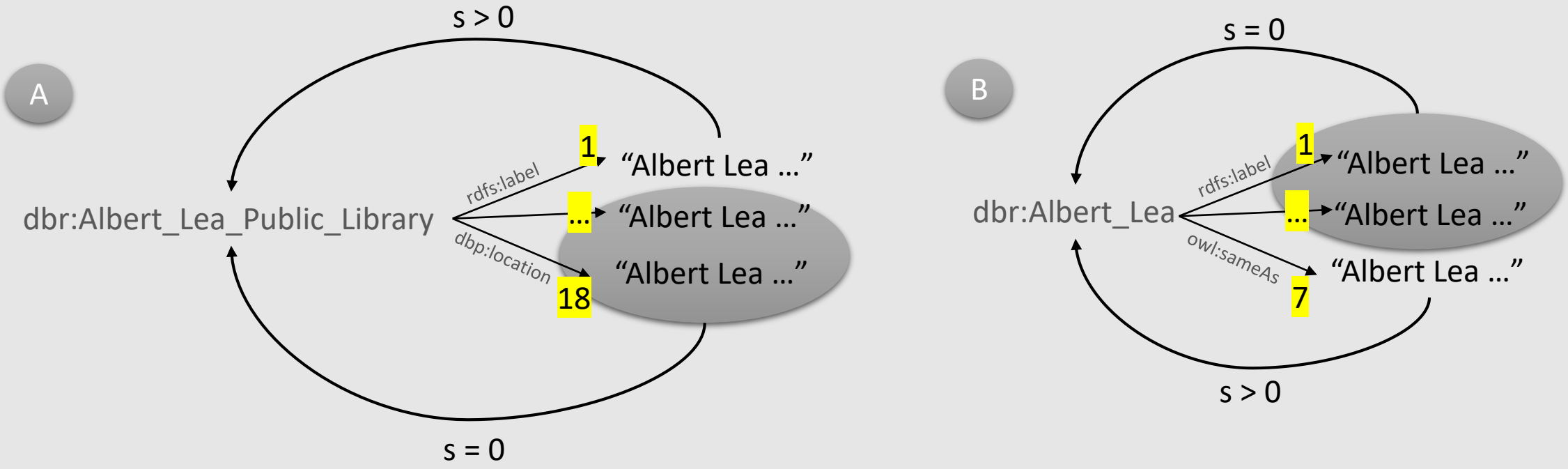
Approach

Can we do better?

Conditional Spread Activation

Challenge 1: Local Term Frequency

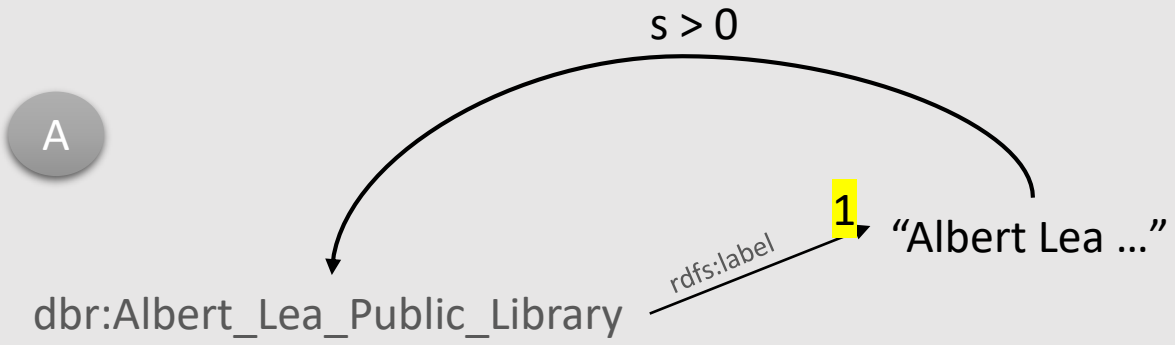
Q: “Albert Lea”

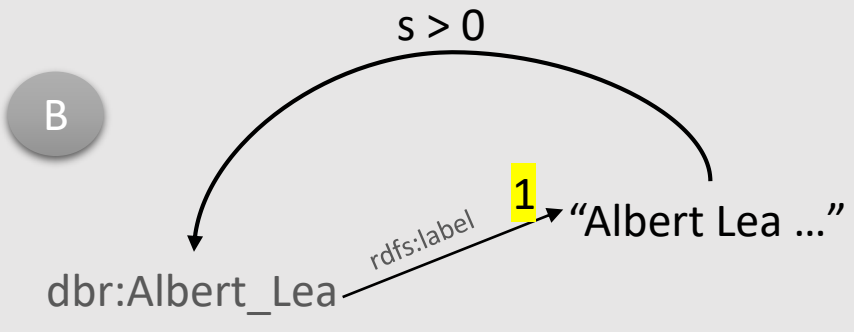


Field Weight

~~Challenge 1: Local Term Frequency~~
Challenge 2: Cohesion

Q: “Albert Lea”





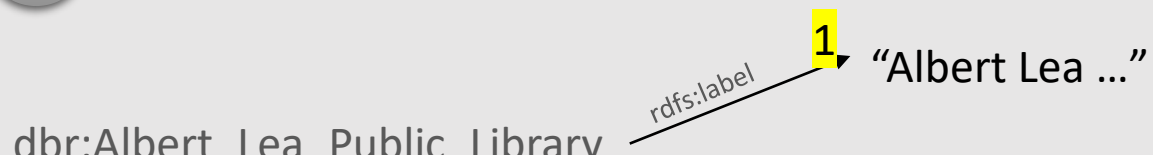
$$weight(p_{type}) > weight(p_{label}) > weight(p_{others})$$

Field Weight

Q: “Albert Lea”

Challenge 1: Local Term Frequency
Challenge 2: Cohesion

A



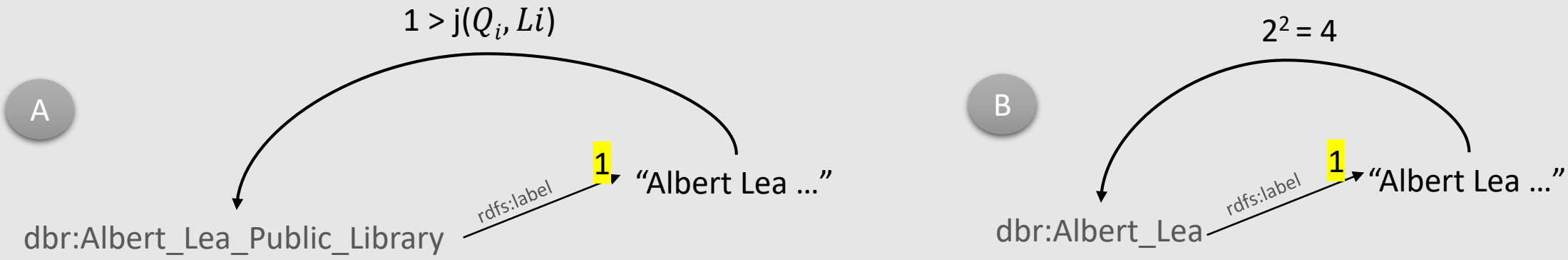
B

$$S(Q_i, L_i) = \sum Q_i^{L_i} \text{ if } \sum Q_i L_i = 1$$

Field Weight

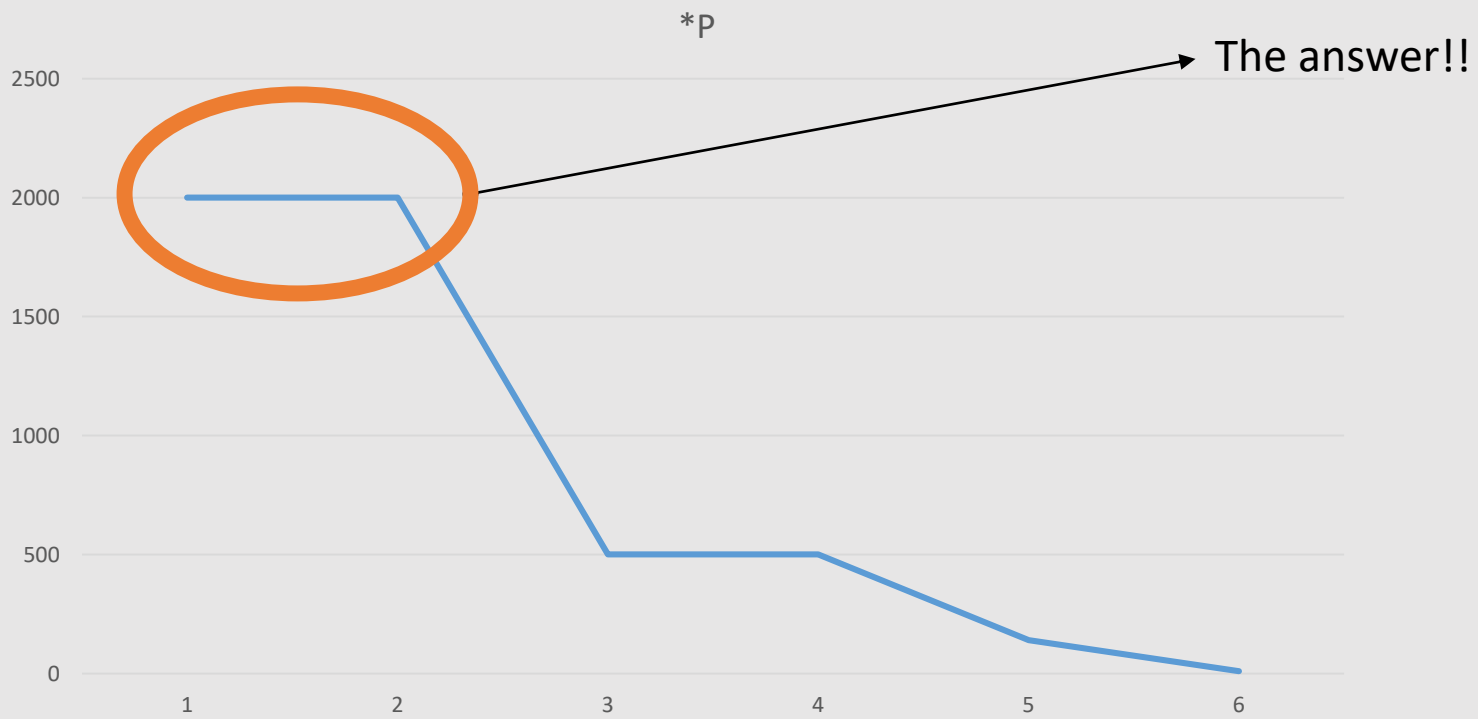
Challenge 1: Local Term Frequency
Challenge 2: Cohesion

Q: “Albert Lea”



$$S(Q_i, L_i) = \sum Q_i^{L_i} \text{ if } \sum Q_i L_i = 1$$

Field Weight



~~Challenge 1: Local Term Frequency~~
~~Challenge 2: Cohesion~~
~~Challenge 3: Scoring Function~~

Pipeline


Query
Query Expansion
Entity Candidate Generation
Ranking
Result

Benchmark

Benchmark	#Queries
QALD-2 (DBpedia-Entity)	140
QALD-4	50

DBpedia-Entity

Approach	MAP
CACAO	**0.2417**
BM25-CA	0.1939
SDM+EL	0.1887
FSDM+EL	0.1719
BM25F-CA	0.1671
LTR	0.1629
LM+EL	0.1534
SDM	0.1533
LM	0.1424
FSDM	0.1403
MLM-CA	0.1273
PRMS	0.1103
BM25	0.1092
MLM-all	0.1062

Table 2: Mean Average Precision (**MAP**) achieved by different Entity Retrieval models on QALD-2 *DBpedia-Entity* benchmark data set.

Approach	P@10
CACAO	**0.3057**
BM25-CA	0.2527
LM+EL	0.2362
SDM+EL	0.2249
LM	0.2144
FSDM+EL	0.2113
BM25F-CA	0.2053
FSDM	0.2000
SDM	0.1883
PRMS	0.1871
MLM-CA	0.1844
MLM-all	0.1843
LTR	0.1732
BM25	0.0986

Table 3: Precision 10 (**P@10**) achieved by different Entity Retrieval models on QALD-2 *DBpedia-Entity* benchmark data set.

QALD-4

Approach	P	R	F_1
CACAO+F	**0.19**	**0.19**	**0.19**
CACAO	0.11	0.11	0.11
CACAO_{P65}	0.09	0.09	0.09
Levenshtein_b	0.04	0.05	0.04
BM25F [2]	0.03	0.03	0.03
Jaccard_b	0.01	0.04	0.01
Levenshtein_a	0.00	0.00	0.00
Jaccard_a	0.00	0.00	0.00

Table 4: ***Precision***, ***Recall*** and $\boldsymbol{F_1}$-*measure* achieved by different ER approaches on QALD-4 benchmark data set.

Approach	P	R	F_1
CACAO_{P65}	**1**	**1**	**1**
CACAO	0.90	0.90	0.90
MAG [24]	0.80	0.80	0.80
DBpedia Spotlight [7]	0.70	0.70	0.70
Levenshtein_b	0.60	0.60	0.60
Jaccard_b	0.60	0.60	0.60
AGDISTIS [36]	0.30	0.30	0.30
BM25F [2]	0.30	0.30	0.30
Levenshtein_a	0.00	0.00	0.00
Jaccard_a	0.00	0.00	0.00

Table 5: ***Precision***, ***Recall*** and $\boldsymbol{F_1}$-*measure* achieved by different EL approaches on QALD 4 benchmark data set.

Conclusion & Future Works

- We have shown a **promising** algorithm that outperform state-of-the-art ER engines on standard benchmarks by addressing:
    - Local Term Frequency;
    - Cohesion, and;
    - Score Function problems
- It is still too soon to trace conclusions (law of small numbers) achieves that
- Improve the algorithm runtime so it can be used at scale

Acknowledgements

@theLiberAI @HTWKLeipzig @edgardmarx @akswgroup

Acknowledgements

Thank you!!!

* 100+ GitHub stars
* 40+ forks

@theLiberAI @HTWKLeipzig @edgardmarx @akswgroup