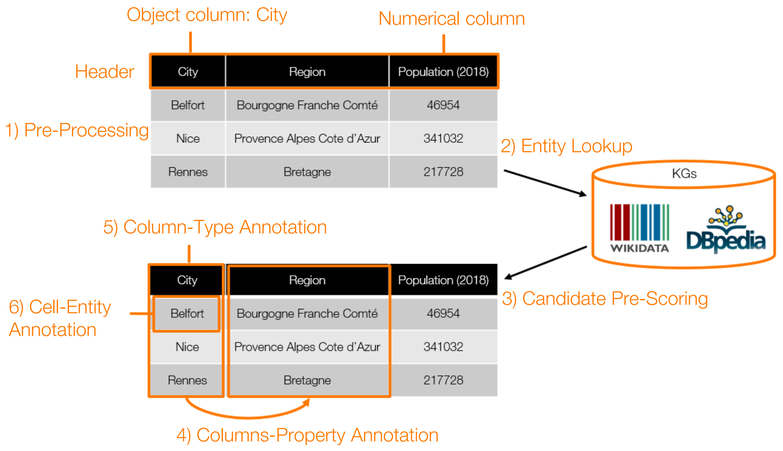
From Heuristics to Language Models
A Journey Through the Universe of Semantic Table Interpretation with DAGOBAH

Viet-Phi Huynh, Yoan Chabot,
Thomas Labbé, Jixiong Liu, Raphaël Troncy
dagobah.support@orange.com

DAGOBAH SL 2022

DAGOBAH Annotation Workflow



Entity Lookup Improvement

Entity	Labels/Aliases (en)
…	…
Q5544925	George Stroumboulopoulos Tonight
Q317521	Elon Musk, Elon Reeve Musk, Elon R. Musk
…	…

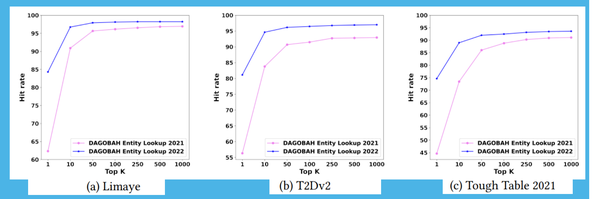
Can **Q5544925** be matched with mention **The Hour** ?
No

Fact: (Wikipedia)
George Stroumboulopoulos Tonight (originally known as **The Hour**) is a Canadian television talk show hosted by George Stroumboulopoulos…

Exploit relevant external alias sources like Wikipedia
Ranking function: incorporate numerous ranking factors (Levenshtein, BM25, PageRank)

Alias source	Wikipedia	Mewsli	Wikilinks	WikiDiverse
Total Number of Wikidata Entity	6 603 252	84 413	1 632 661	7704
Number of Wikidata Entity whose aliase set is enriched	5 537 830	26 989	992 888	2831
Average number of novel aliases enriched per entity	2.5	1.5	3.1	1.2

HitRate@K of new and old entity lookup



Entity Scoring Improvement

Score of a candidate entity $\hat{e}$ of m :
E.g. m = **LAGUARDIA NY,US AIRWAYS*** $\hat{e}$ = **Q319654 (Laguardia Airport)**
$p(\hat{e} =$ **Q319654**$) = ?$

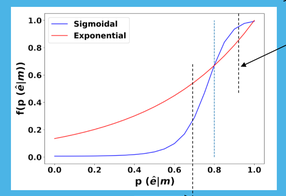
$$p(\hat{e}) = p(\hat{e}|\text{table context}) \times f(p(\hat{e}|m))$$

activation function

p (**Q319654** | **Airplane, > 1000 ft, LakeFront Airport**…) calculated by DAGOBAH SL 2021 [*]

prior p(**Q319654** | **LAGUARDIA NY US AIRWAYS***) resulted from entity lookup

The choice of activation function f is important



p > 0.9: more likely good candidate

With **sigmoidal** f,
{p>0.9} is better distinguished from **{p<0.7}**

Function f	Valid HardTable R2	Valid ToughTable R2
Exponential	0.888	0.941
Sigmoidal	0.907	0.959

F1 score on Validation datasets of Round 2

p < 0.7: more likely bad candidate

Entity Disambiguation by Reading Entity Descriptions

Table dffeec8c3593402bafa69b50f5920fa5.csv (BioDivTable)

aircraft type	airport name	altitude bin	…
…	…		…
Airplane	LAGUARDIA NY,US AIRWAYS*	> 1000 ft	…
…	LAKEFRONT AIRPORT,BUSINESS	…	…

Column headers, if appropriately given, can help to disambiguate the entity
Not Laguardia (Spain municipality)
Not US AIRWAYS (Airline Company)
But LaGuardia Airport (header **airport name**)

How to evaluate if **LaGuardia Airport** (and not **Laguardia municipality, US AIRWAYS**) is relevant w.r.t. headers H = **[aircraft type, airport name, altitude bin…]?**
→ **read their descriptions d_e**

Modelling f : ELECTRA-based Cross Encoder [*] fine tuned on Wikipedia Table

f ([aircraft type, **airport name**, altitude bin], **LaGuardia Airport (IATA: LGA, ICAO: KLGA, FAA LID: LGA) is a civil airport in East Elmhurst, Queens, New York City…**) = **0.99**

f ([aircraft type, **airport name**, altitude bin], **US Airways (formerly USAir) was a major U.S. airline that operated from 1937 until its merger with American Airlines in 2015…**) = **0.12**

f ([aircraft type, **airport name**, altitude bin], **Laguardia (Basque: Guardia) is a town and municipality located in the southern province of Álava, in the north of Spain; it belongs to the region of Rioja Alavesa…**) = **0.009**

Results

SemTab @ ISWC 2022 Results

Dataset	System	CTA		CEA		CPA	
		F1	P	F1	P	F1	P
Round 1 - Hard Table WD	**DAGOBAH SL**	**0.975**	**0.975**	**0.954**	**0.955**	**0.984**	**0.99**
Round 2 - Hard Table WD	**DAGOBAH SL**	**0.96**	**0.96**	**0.904**	**0.905**	**0.931**	**0.97**
Round 2 – Tough Table WD	**DAGOBAH SL**	**0.409**	**0.409**	**0.945**	**0.946**	**-**	**-**
Round 2 – Tough Table DBP	**DAGOBAH SL**	**0.312**	**0.312**	**0.926**	**0.926**	**-**	**-**
Round 3 – BioDivTable	**DAGOBAH SL + Header Disambiguation**	**0.616**	**0.616**	**0.736**	**0.736**	**-**	**-**
Round 3 - GitTable	**DAGOBAH SL**	**0.075**	**0.082**	**0.312**	**0.342**	**0.087**	**0.095**

Conclusion and Future Work

DAGOBAH SL 2022 in a nutshell
- Entity lookup and scoring improvement
- Entity disambiguation by reading entity descriptions
- Performance optimization (~30% gain in processing time)

Leverage Deep Learning methods to deal with table complexity
- Use graph embeddings to deal with the most ambiguous cases: see “Radar Station”
- Use Language Model + Cross encoder to better understand the headers of a table
- Consider table as a “language” to leverage richer representations

From Wikidata to Enterprise Knowledge Graphs: Use enterprise knowledge graphs to annotate business-related data instead of Wikidata where entities are less richly described

Test our tools!

DAGOBAH API
https://developer.orange.com/apis/table-annotation
(for logged in users)

DAGOBAH UI Demo
https://tinyurl.com/dagobah-ui-demo

Radar Station
https://github.com/Orange-OpenSource/radar-station