Constrained Information Retrieval for Long-Tail Knowledge Extraction

Nicolas Lazzari

University of Pisa
University of Bologna

nicolas.lazzari3@unibo.it

Arianna Graciotti

University of Bologna

arianna.graciotti@unibo.it

Valentina Presutti

University of Bologna

valentina.presutti@unibo.it

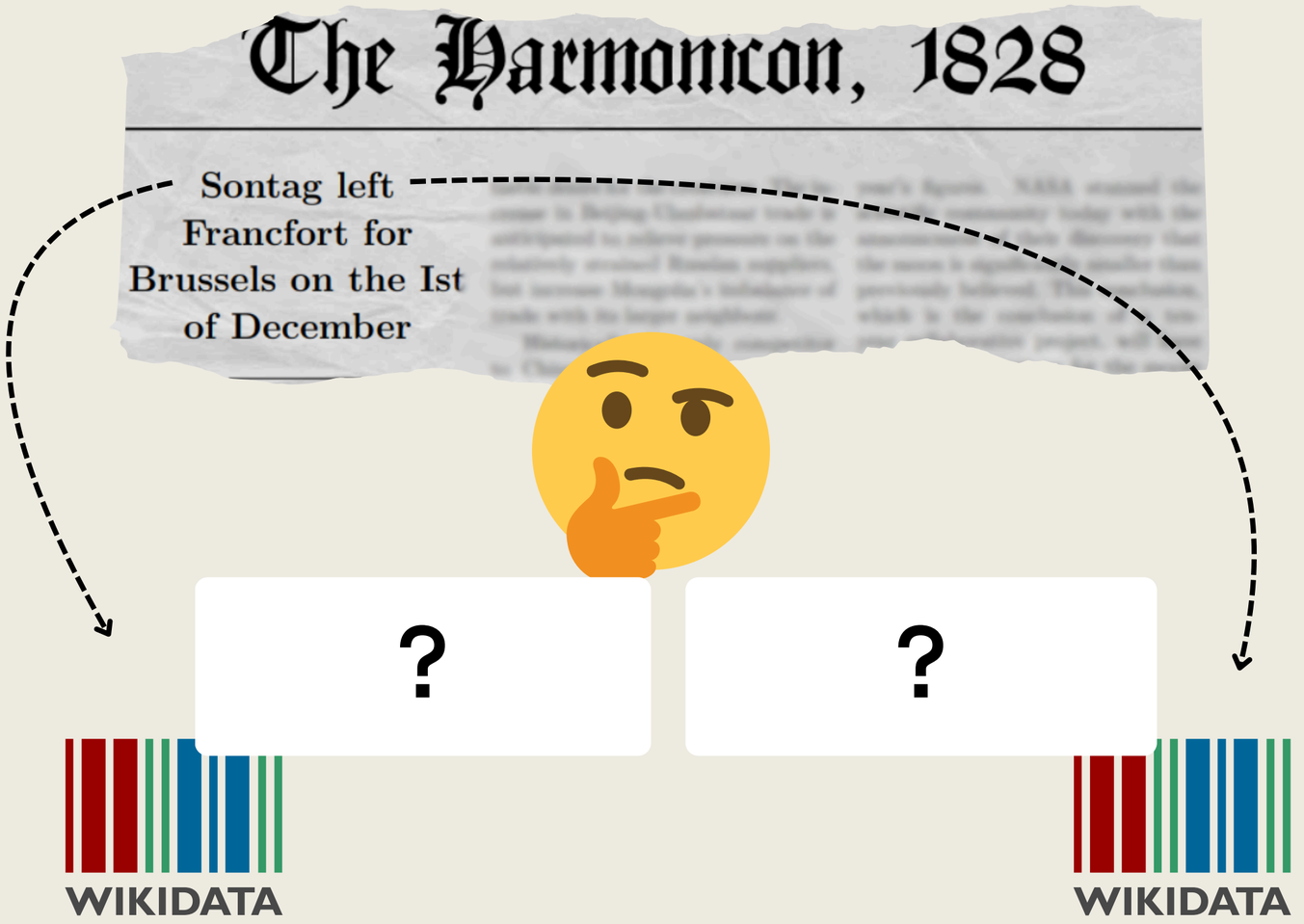
Future Artificial Intelligence Research

Polifonia

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

UNIVERSITÀ DI PISA

MOTIVATION

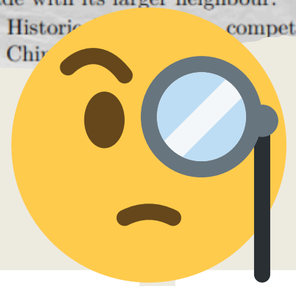

The Harmonicon, 1828
Sontag left
Francfort for
Brussels on the Ist
of December
?
?
WIKIDATA
WIKIDATA


The Harmonicon, 1828

Sontag left Francfort for Brussels on the Ist of December

crease in Beijing-Ulanbataar trade is anticipated to relieve pressure on the relatively strained Russian suppliers, but increase Mongolia's imbalance of trade with its larger neighbour.

year's figures. NASA stunned the announcement of their discovery that the moon is significantly smaller than previously believed. This conclusion, which is the conclusion of a ten-



Susan Sontag - Q152824

writer

Henriette Sontag - Q64098

operatic singer

Susan Sontag - Q152824
1933-2004
writer

Henriette Sontag - Q64098
1806-1854
operatic singer

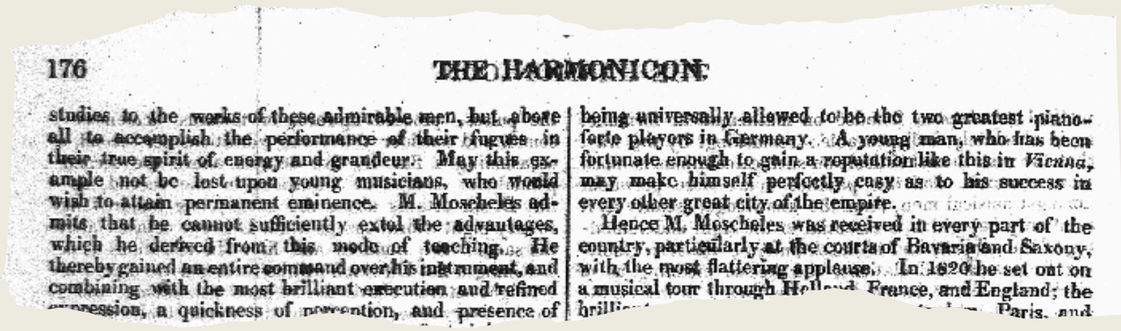
Historical documents: a blind spot for (L)LMs

NLP research is primarily focused on **contemporary**, well-edited text, mostly in English

...etc.

Historical documents: a blind spot for (L)LMs

NLP research is primarily focused on **contemporary**, well-edited text, mostly in English


176 THE HARMONICON

studies to the works of these admirable men, but above all to accomplish the performance of their fugues in their true spirit of energy and grandeur. May this example not be lost upon young musicians, who would wish to attain permanent eminence. M. Moscheles admits that he cannot sufficiently extol the advantages, which he derived from this mode of teaching. He thereby gained an entire command over his instrument, and combining with the most brilliant execution and refined expression, a quickness of perception, and presence of

being universally allowed to be the two greatest pianoforte players in Germany. A young man, who has been fortunate enough to gain a reputation like this in *Vienna*, may make himself perfectly easy as to his success in every other great city of the empire.

Hence M. Moscheles was received in every part of the country, particularly at the courts of Bavaria and Saxony, with the most flattering applause. In 1820 he set out on a musical tour through Holland, France, and England; the

Paris, and

Historical documents pose unique challenges:

- OCR noise,
- Language variations,

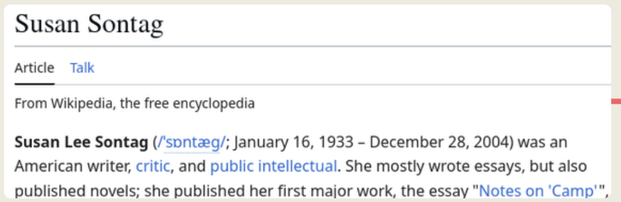
Historical documents: a blind spot for (L)LMs

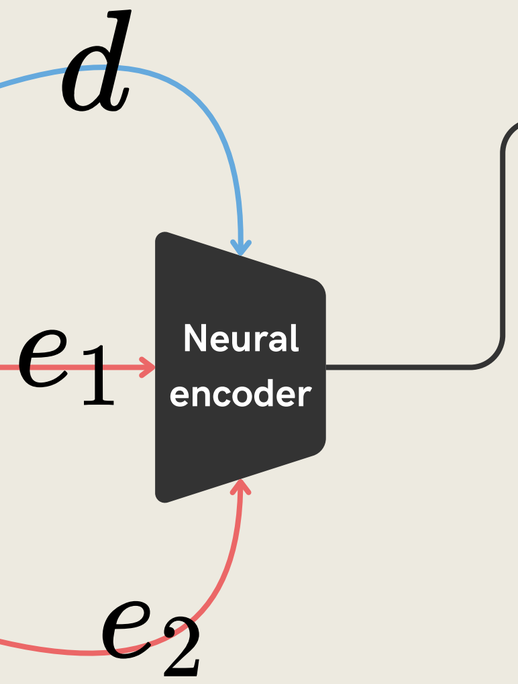
NLP research is primarily focused on **contemporary**, well-edited text, mostly in English

Historical documents pose unique challenges:

- OCR noise,
- Language variations,
- Lesser-known (long-tail) entities

Historical documents: a blind spot for (L)LMs <u>because of representation</u>



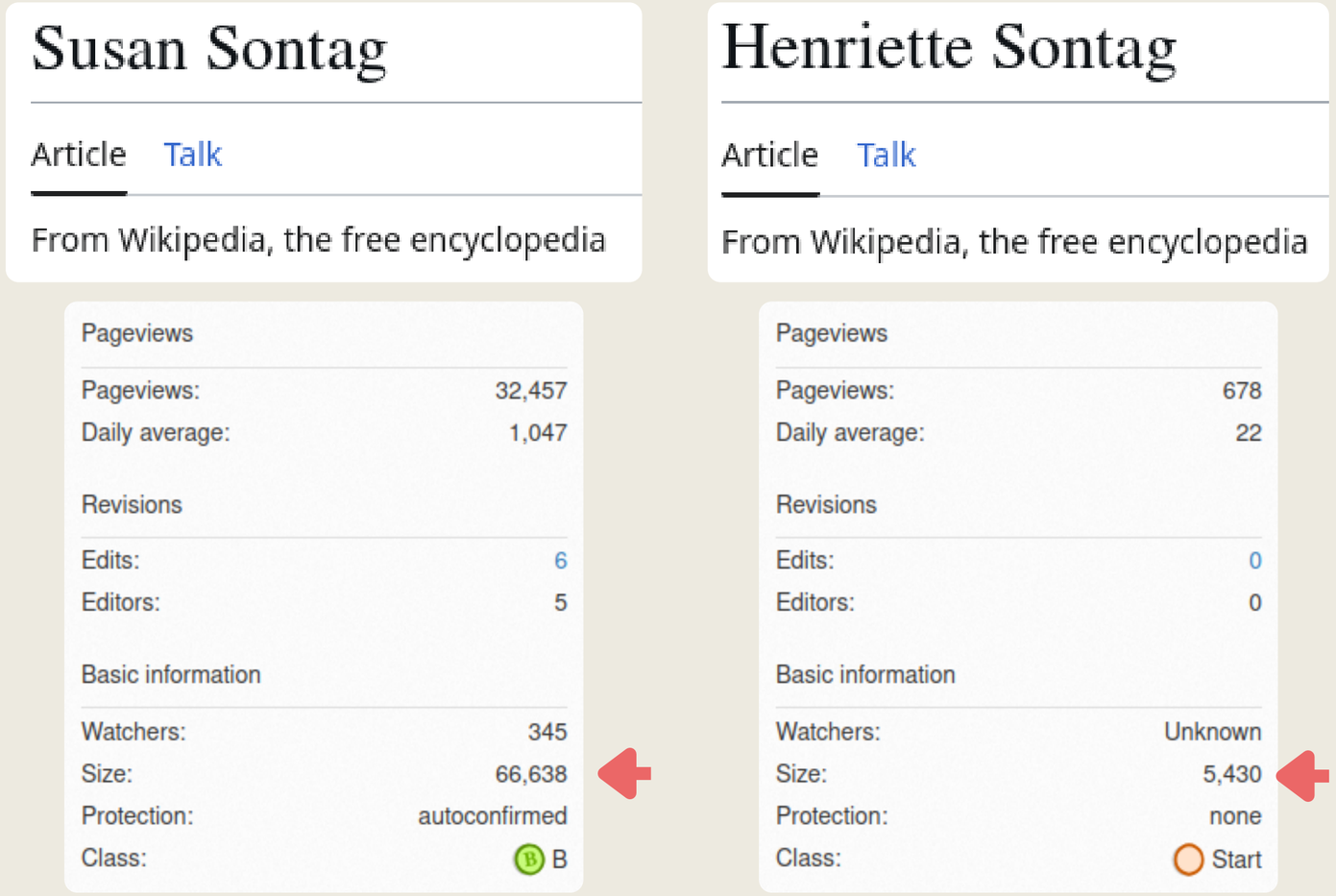


$$\vec{d} = [0.3, -0.23, \cdots]$$
$$\vec{e_1} = [-0.42, 0.11, \cdots]$$
$$\vec{e_2} = [0.03, 0.01, \cdots]$$

Similarity computed as cosine distance **heavily relies** on the accuracy of entity representations

$$\langle \vec{d}, \vec{e_1} \rangle > \langle \vec{d}, \vec{e_2} \rangle$$

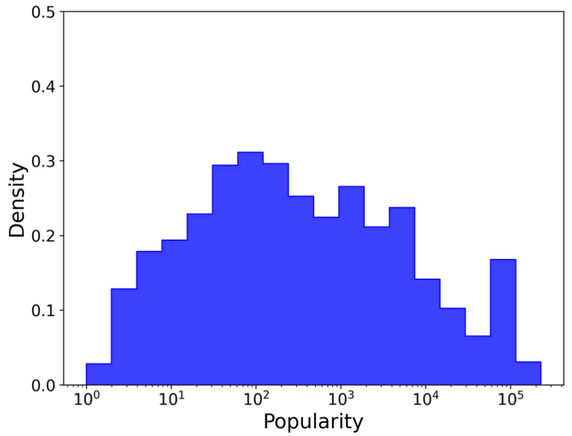
Historical documents: a blind spot for (L)LMs because of popularity



Information retrieval methods consistently exploit Wikipedia for **entity descriptions**

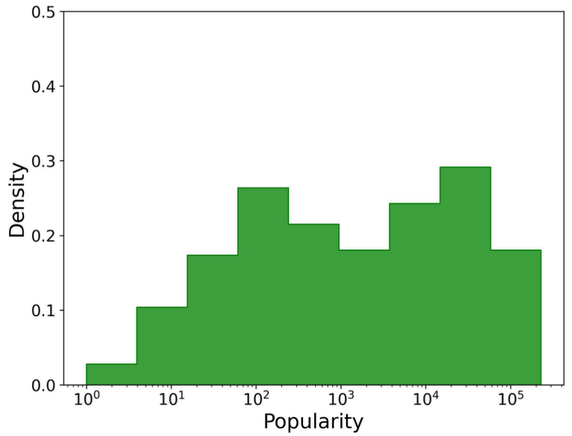
More popular entities have longer descriptions and result in **less shallow representations**

Historical documents are especially long-tail benchmarks



MHERCL

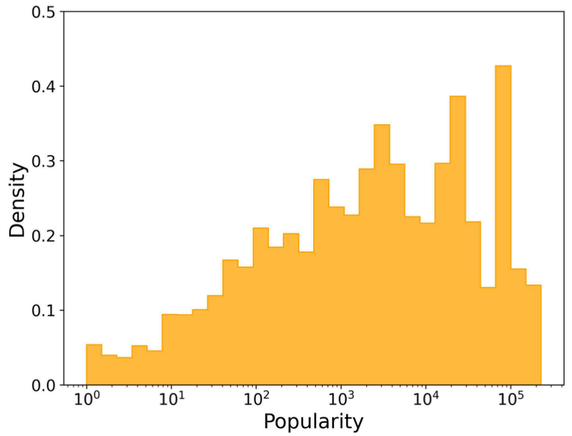
Entities popularity distribution



HIPE-2020

Entities popularity distribution

Historical newspapers benchmarks



AIDA CONLL-YAGO

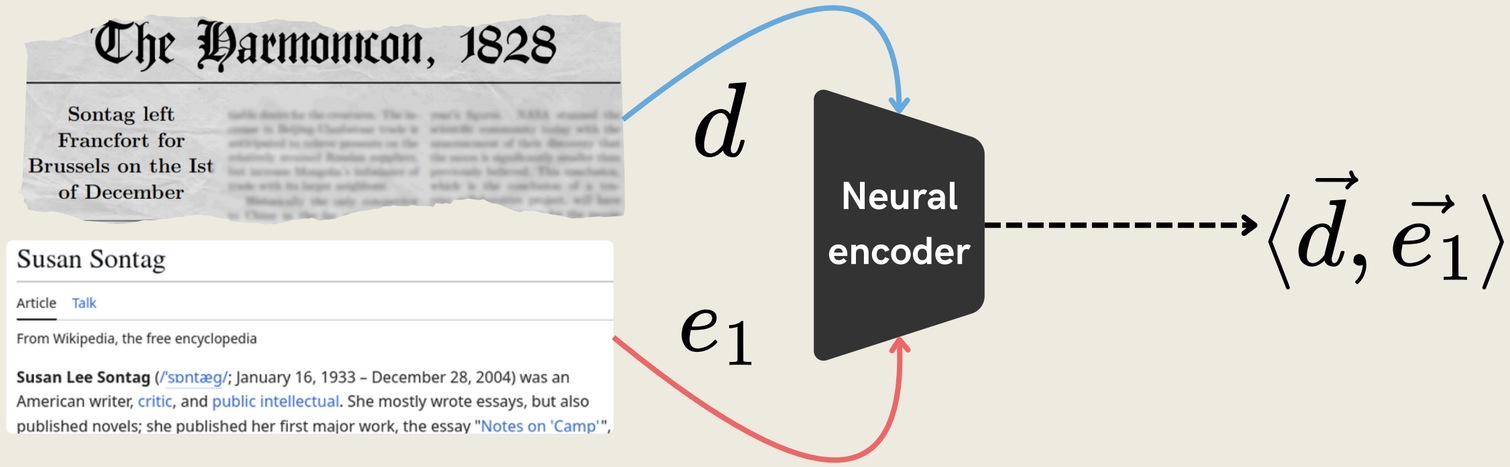
Entities popularity distribution

Nowadays news benchmark

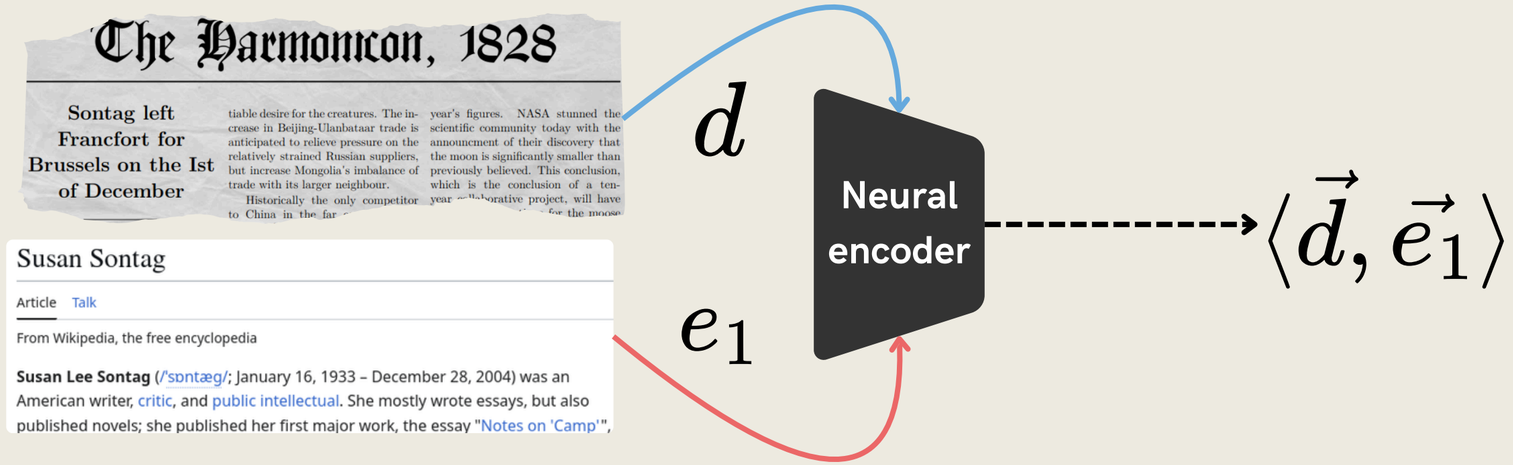
RESEARCH QUESTIONS

[RQ1] What challenges affect the **retrieval** of unpopular entities?

[RQ2] How can we **enhance** (L)LMs' performance in retrieving these entities?



INTUITION



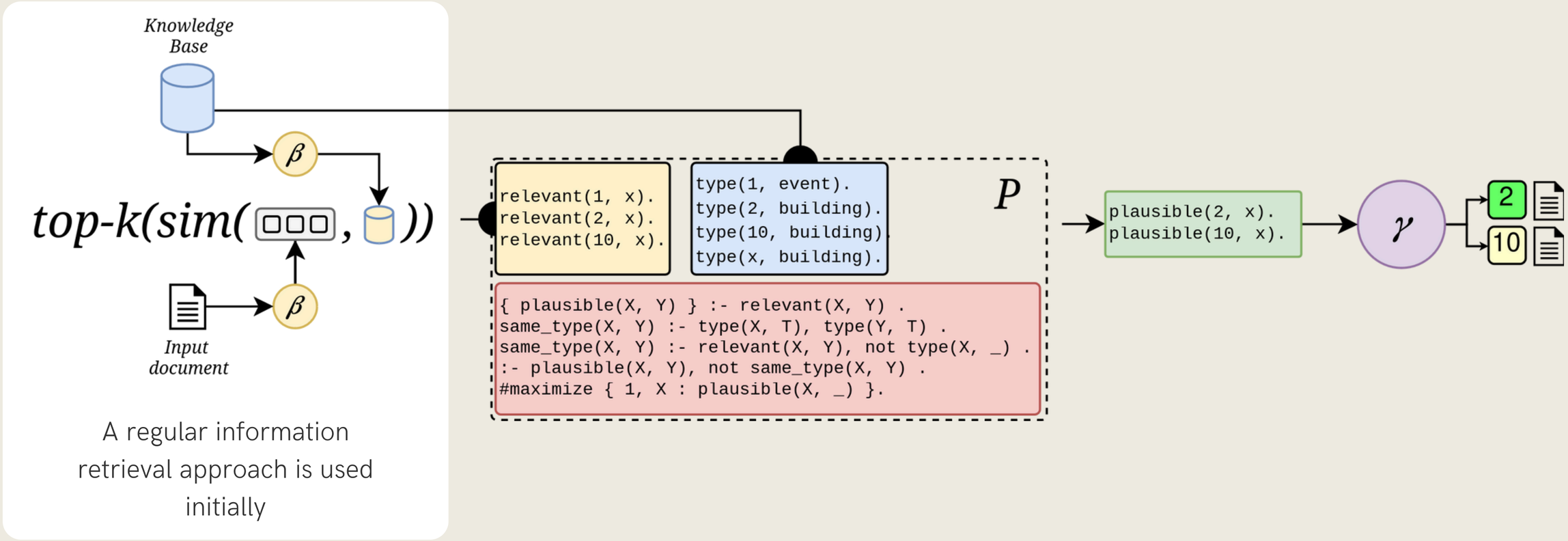
...Why even bother checking against implausible entities?

CONTRIBUTIONS

- A method based on **Answer Set Programming** (**ASP**) that imposes logical plausibility **constraints** on the output of LM-based retrieval systems.
- Tests on four **historical documents benchmarks** annotated for the **Entity Linking** task show our method **boosts recall** and surpasses specialized models.

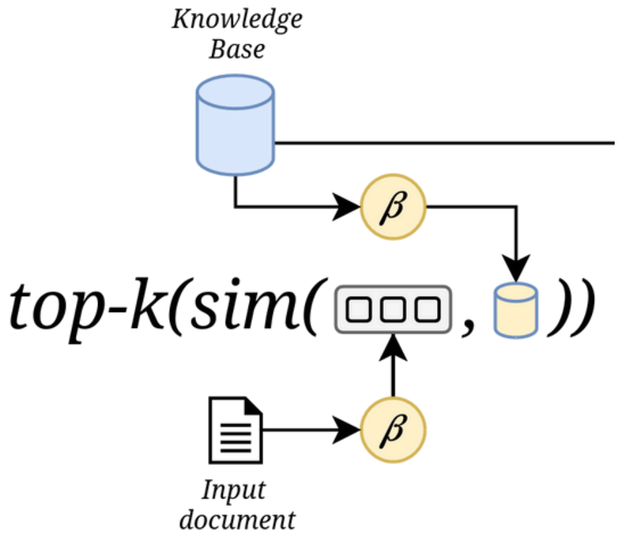
PROPOSED METHOD

Constraining information retrieval through Answer Set Programming



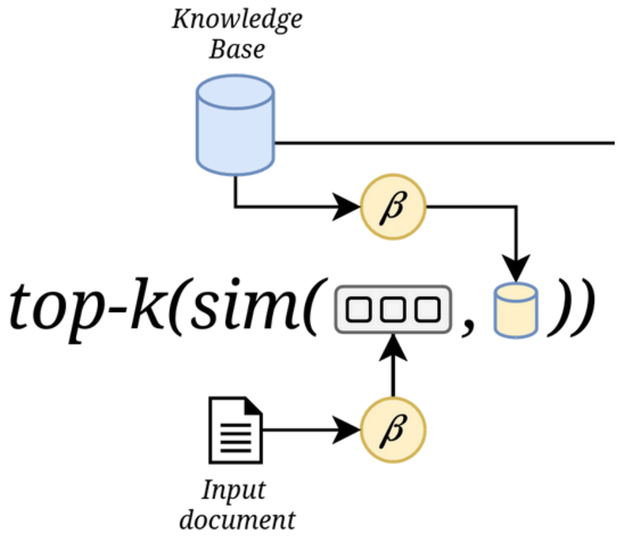
A regular information retrieval approach is used initially

Constraining information retrieval through Answer Set Programming



We exploit datasets annotated for **entity linking** and interpret an annotated named entity as a the retrieval **query**

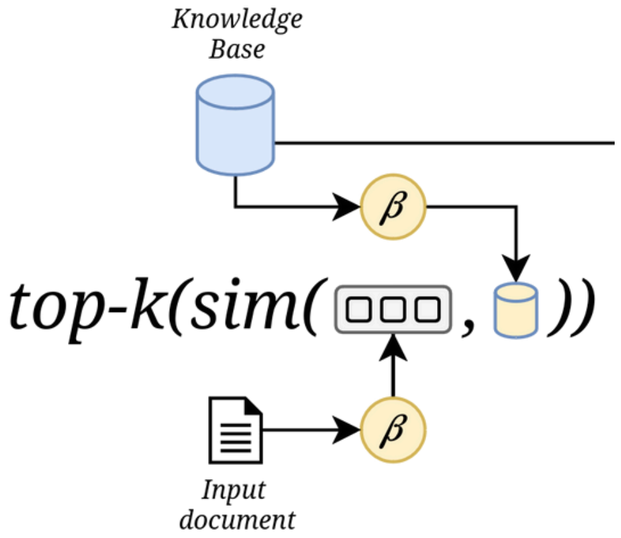
Constraining information retrieval through Answer Set Programming



We exploit datasets annotated for **entity linking** and interpret an annotated named entity as a the retrieval **query**

The document encoder β is a **regular sentence embedding method** (MPNet, distill-RoBERTa, MiniLM)

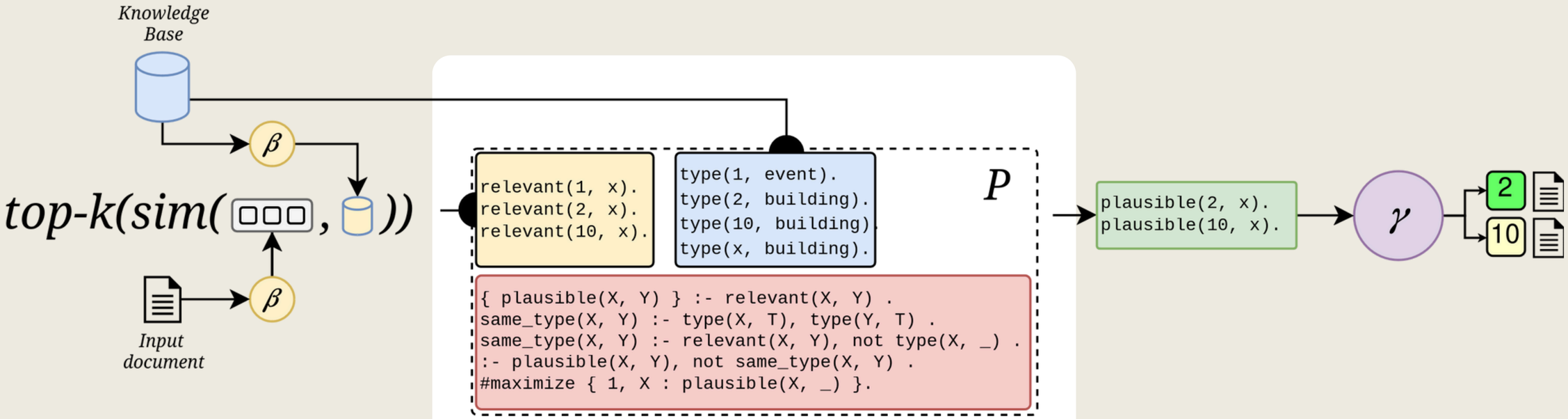
Constraining information retrieval through Answer Set Programming



We encode the input sentence using β and **bias** it towards the input entity by projecting it on the embedding of the entity computed with β

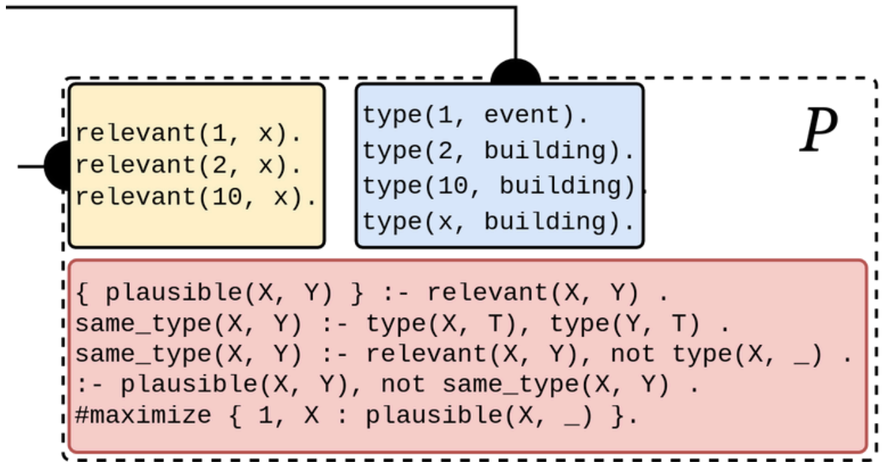
The encoder β is not finetuned for entity retrieval, hence it is not biased because of standard datasets.

Constraining information retrieval through Answer Set Programming



A logic program in **ASP** formalizes the logical plausibility constraints.

Constraining information retrieval through Answer Set Programming



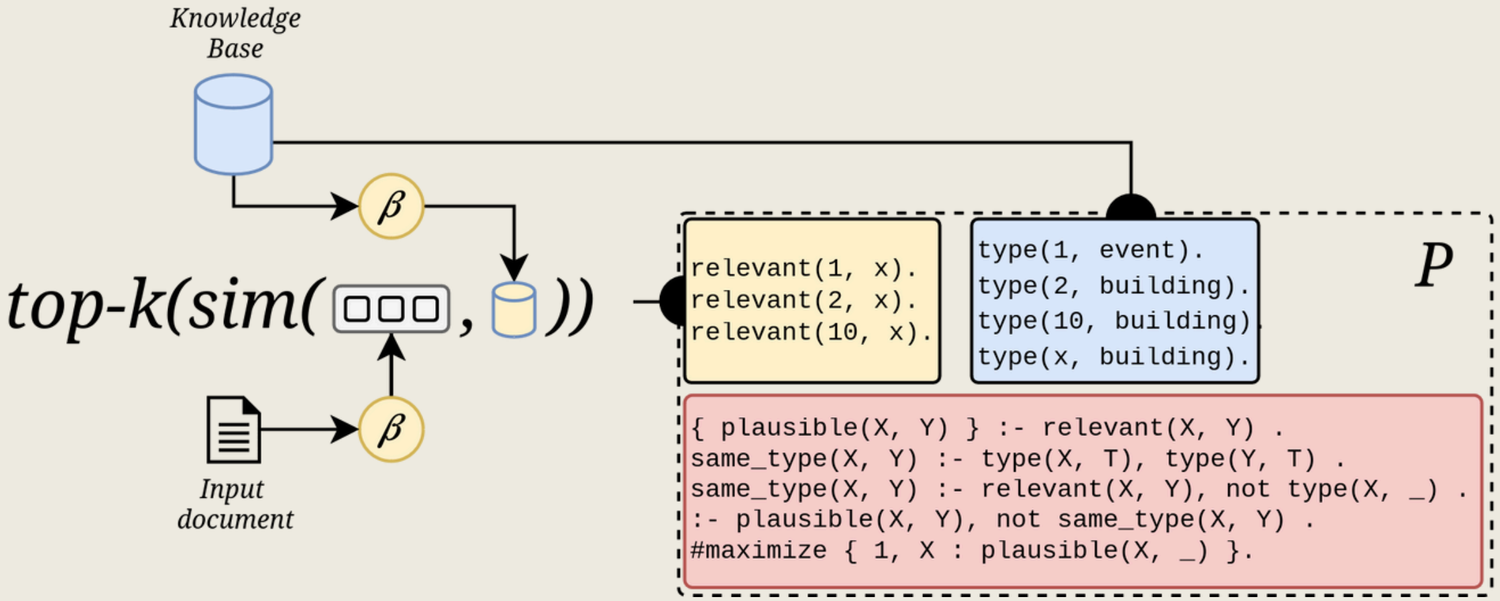
A logic program in **ASP** formalizes the logical plausibility constraints.

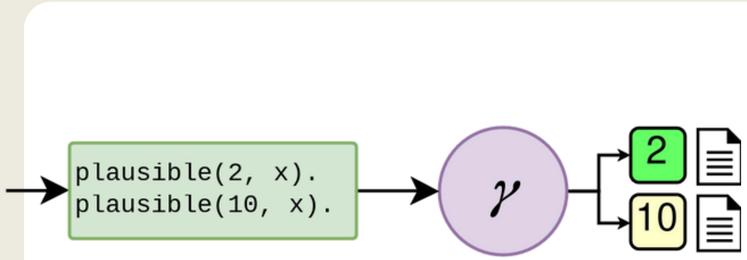
```
% Generate plausible candidates
{ plausible(X, Y) } :- relevant(X, Y) .
% Define type-plausibility and remove implausible candidates
same_type(X, Y) :- type(X, T), type(Y, T) .
same_type(X, Y) :- relevant(X, Y), not type(X, _) .
:- plausible(X, Y), not same_type(X, Y) .
% Define year-plausibility and remove implausible candidates
compatible_year(X, Y) :- year(X, YX), year(Y, YY), YX <= YY .
compatible_year(X, Y) :- relevant(X, Y), not year(X, _) .
:- plausible(X, Y), not compatible_year(X, Y) .
% Compute the answer set with the highest number of plausible candidates
#maximize { 1, X : plausible(X, _) }.
```

Type plausibility: a plausible entity must be classified with the same type of the named entity.

Date plausibility: a plausible entity must be have an associated Wikidata date that precedes the one of the input document.

Constraining information retrieval through Answer Set Programming





The stable answer set computed on the ASP program is the set of **plausible** and **relevant** documents given the input document and its query.

RESULTS

HIPE2020 ***(Annotations: Entity Linking)***

a dataset of 19C US historical newspapers

Model		R@10	R@30	R@50	R@100	R@200	R@300
ReLiK [24]		**0.81**	0.90	0.93	0.96	0.97	1.00
MPNet [35]		0.42	0.62	0.73	0.89	0.99	1.00
	+ ASP	0.65	**0.91**	**0.96**	**0.99**	0.99	1.00
distill-RoBERTa [36]		0.39	0.58	0.71	0.83	0.94	1.00
	+ ASP	0.59	0.87	0.94	**0.99**	**1.00**	1.00
MiniLM [37]		0.31	0.49	0.59	0.82	0.96	1.00
	+ ASP	0.51	0.84	0.95	**0.99**	**1.00**	1.00

Type plausibility **Date plausibility**

HIPE2020 *(Annotations: Entity Linking)*

a dataset of 19C US historical newspapers

Model		R@10	R@30	R@50	R@100	R@200	R@300
ReLiK [24]		**0.81**	0.90	0.93	0.96	0.97	1.00
MPNet [35]		0.42	0.62	0.73	0.89	0.99	1.00
	+ ASP	0.65	**0.91**	**0.96**	**0.99**	0.99	1.00
distill-RoBERTa [36]		0.39	0.58	0.71	0.83	0.94	1.00
	+ ASP	0.59	0.87	0.94	**0.99**	**1.00**	1.00
MiniLM [37]		0.31	0.49	0.59	0.82	0.96	1.00
	+ ASP	0.51	0.84	0.95	**0.99**	**1.00**	1.00

Type plausibility

Date plausibility

MHERCL *(Annotations: Entity Linking)*

a dataset of British music magazines of the 19C

Model		R@10	R@30	R@50	R@100	R@200	R@300
ReLiK [24]		**0.84**	0.91	0.93	0.96	0.99	1.00
MPNet [35]		0.38	0.65	0.73	0.88	0.97	1.00
	+ ASP	0.72	**0.92**	**0.96**	**0.99**	**1.00**	1.00
distill-RoBERTa [36]		0.39	0.58	0.71	0.82	0.96	1.00
	+ ASP	0.68	0.87	**0.96**	**0.99**	**1.00**	1.00
MiniLM [37]		0.27	0.49	0.61	0.78	0.92	1.00
	+ ASP	0.68	0.89	0.95	**0.99**	**1.00**	1.00

Type plausibility

Date plausibility

AjMC *(Annotations: Entity Linking)*

a dataset of 19C commentaries about Sophocle's tragedy "Ajax"

Model		R@10	R@30	R@50	R@100	R@200	R@300
ReLiK [24]		**0.90**	0.93	0.93	0.94	0.99	1.00
MPNet [35]		0.38	0.50	0.52	0.94	0.98	1.00
	+ ASP	0.51	**0.96**	**1.00**	**1.00**	**1.00**	1.00
distill-RoBERTa [36]		0.29	0.47	0.51	0.92	**1.00**	1.00
	+ ASP	0.45	0.95	**1.00**	**1.00**	**1.00**	1.00
MiniLM [37]		0.23	0.39	0.39	0.50	0.98	1.00
	+ ASP	0.39	0.51	0.98	**1.00**	**1.00**	1.00

Type plausibility

Date plausibility

TopRes19th *(Annotations: Entity Linking)*

a dataset of 18C-19C British library documents (scope restricted to toponyms)

Model		R@10	R@30	R@50	R@100	R@200	R@300
ReLiK [24]		**0.83**	0.90	0.91	0.93	0.97	1.00
MPNet [35]		0.30	0.64	0.76	0.87	0.98	1.00
	+ ASP	0.73	**0.98**	0.99	**1.00**	**1.00**	1.00
distill-RoBERTa [36]		0.34	0.56	0.71	0.85	0.94	1.00
	+ ASP	0.59	0.72	0.99	**1.00**	**1.00**	1.00
MiniLM [37]		0.21	0.42	0.61	0.76	0.95	1.00
	+ ASP	0.62	0.95	**1.00**	**1.00**	**1.00**	1.00

Type plausibility **Date plausibility**

MHERCL

	Model	Top 10
		The Harmonicon, 1828
		Sontag (*Q64098*) left Francfort for Brussels on the Ist of December.
✗	Relik	Brussels [Q240], Sontag [Q47519541], Alan Sontag [Q945286], Susan Sontag [Q152824], Belfort [Q171545], ...
✗	MPNet	Sontag [Q47519541], Sontag, MS [Q7562392], Sonbolabad [Q7560867], Sondor (disambiguation) [Q22349595], Frank Sontag [Q5489708], ...
✓	MPNet + ASP	Sontag [Q47519541], Soner [Q962275], **Henriette Sontag [Q64098]**, Sonam [Q7560775], Ernst Sonntag [Q19661367], ...

Type plausibility **Date plausibility**

TopRes19th

TopRes19th, 1863

And that an AUDIT for the RESERVED and CHIEF RENTS for the Manor of Stayley, in the county of Chester (*Q23064*), will be holden at the Eagle Inn, in Stalybridge, on Thursday, the 7th day of May next, between the hours of Eleven and Two o clock, on which days the tenants are requested to pay their rents.

	Model	Top 10
✗	Relik	Chester [Q170263], Justice of Chester [Q616310], Earl of Chester [Q1277249], Earl of Warrington [Q5326386], Exchequer of Chester [Q5419617], ...
✗	MPNet	Chester County [Q227112], Chester County Courthouse [Q1070703], Chester County History Center [Q19866503], New Chester [Q16462307], Diocese of Chester [Q543301], ...
✓	MPNet + ASP	Chester Rural District [Q5093705], 1724 Chester Courthouse [Q4552563], Chester County, Pennsylvania [Q27840], Chester (town), Orange County, New York [Q2756901], **Cheshire [Q23064]**, ...

From:

Cheshire's name was originally derived from an early name for Chester,

From: https://en.wikipedia.org/wiki/Cheshire

Type plausibility

Date plausibility

Conclusions:
A little semantics goes a long tail!

- **Information Retrieval** empowers a lot of applications (EL, RAG, etc.) and it **can greatly benefit from logical constraints**
- **ASP** is a highly scalable, intuitive and convenient technology to achieve **neuro-symbolic integrations**
- A **simple sentence embedding** method **+ ASP** might be **more than enough** to retrieve your data!

THANKS!

Nicolas Lazzari

University of Pisa
University of Bologna

nicolas.lazzari3@unibo.it

Arianna Graciotti

University of Bologna

arianna.graciotti@unibo.it

Valentina Presutti

University of Bologna

valentina.presutti@unibo.it

Future Artificial Intelligence Research

Polifonia

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

UNIVERSITÀ DI PISA