Using Knowledge Graphs for Text Retrieval

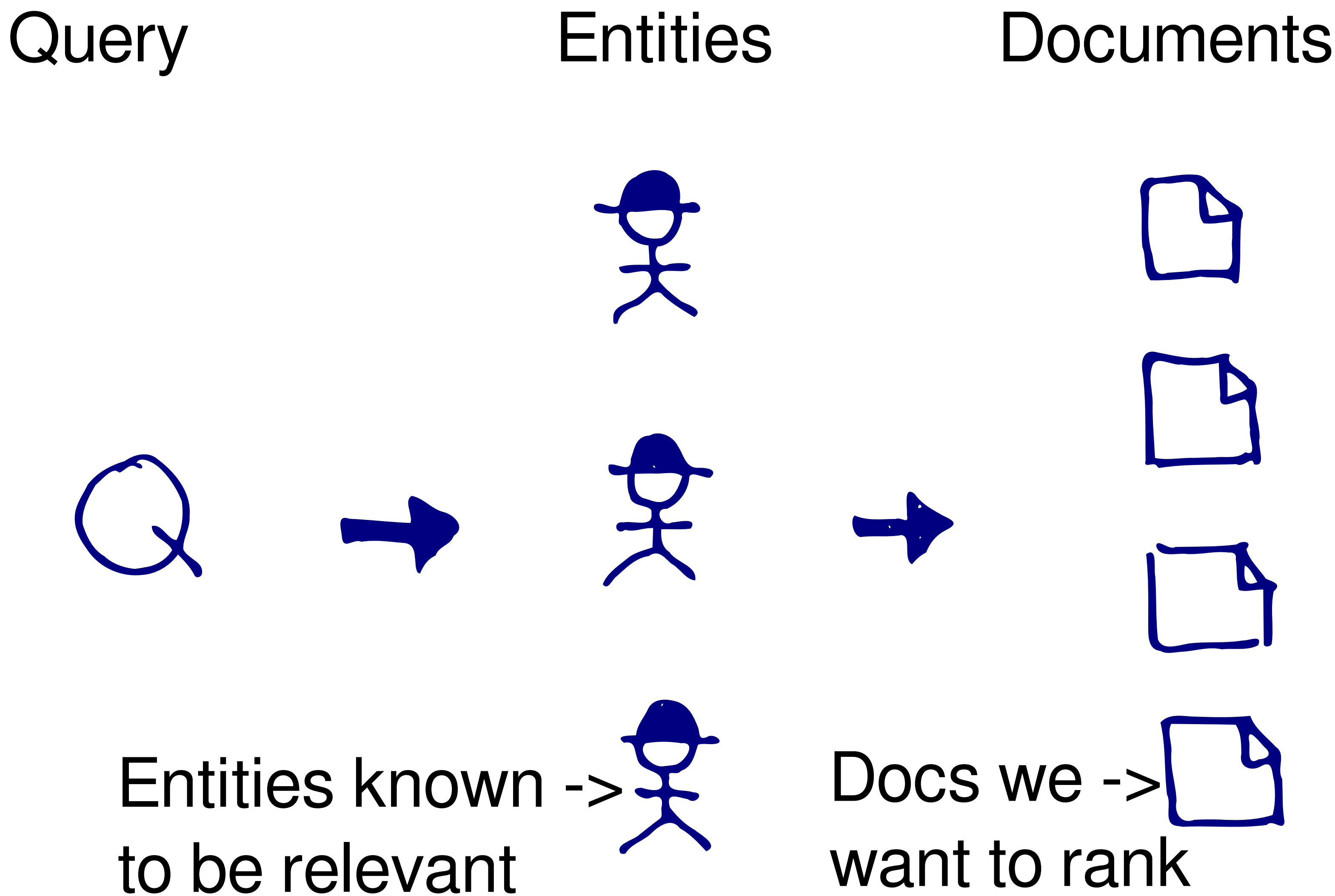
github.com/laura-dietz/tutorial-utilizing-kg

Laura Dietz
University of New Hampshire

Alex Kotov
Wayne State University

Edgar Meij
Bloomberg

Document Retrieval with Entities



Outline

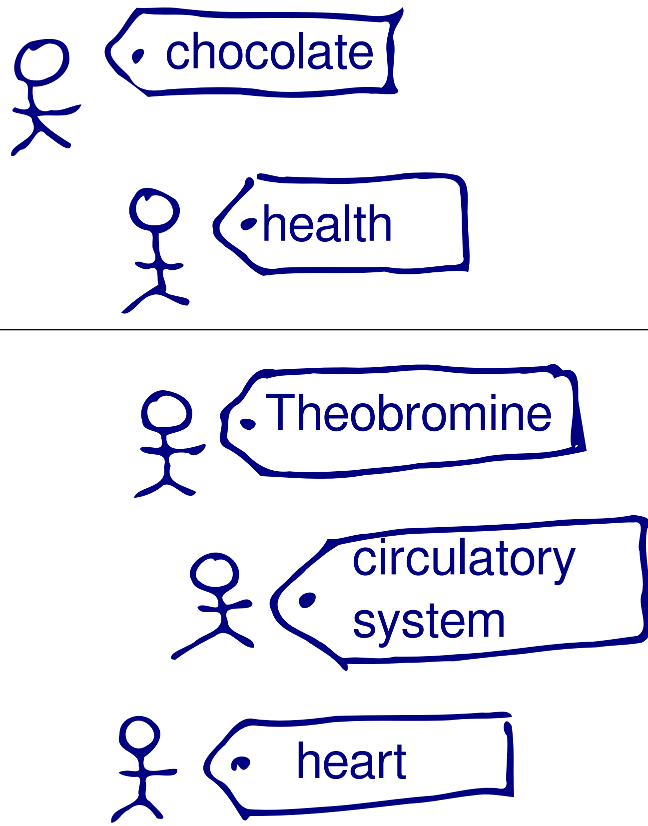
1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

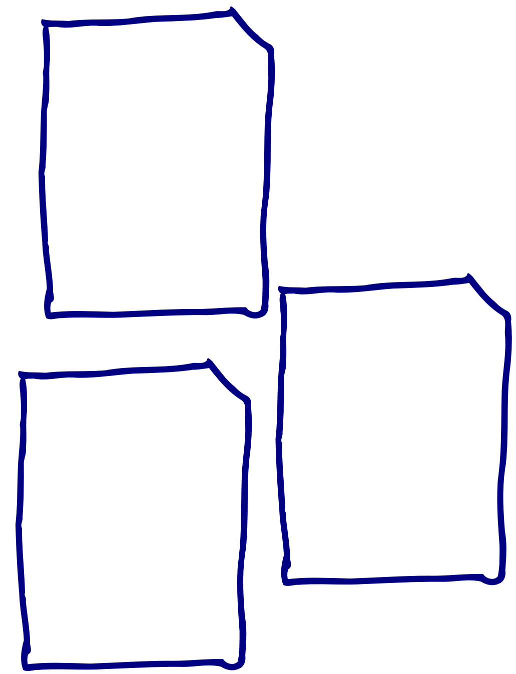
Different Queries - Different Entities

Query	EU UK relations	dark chocolate health benefits
Query entities	EU UK	chocolate health
Latent entities	Brexit Theresa May	Theobromine circulatory system heart
[Hasibi ICTIR16]	**Named Entities**	**Concepts**

Matching Entities in Documents

Q: dark chocolate health benefits

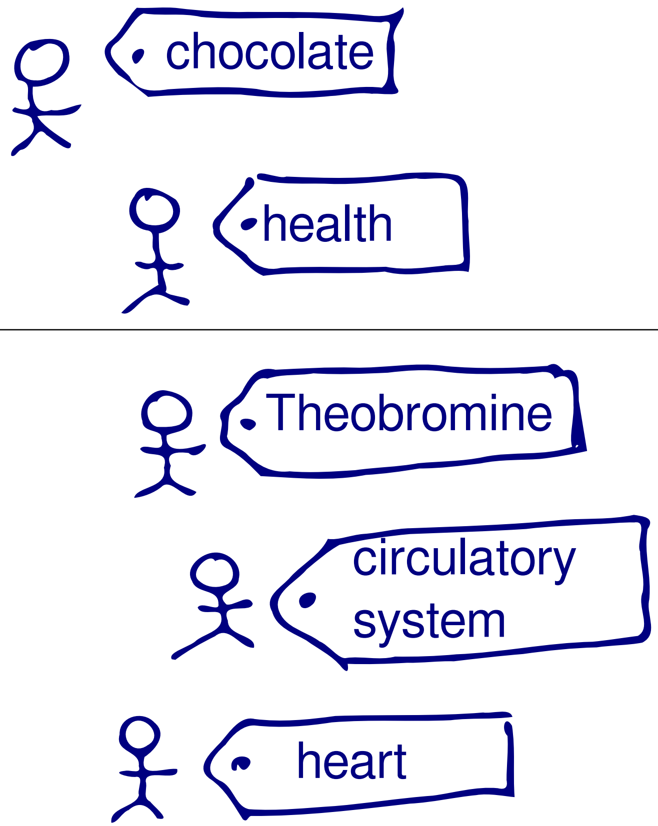


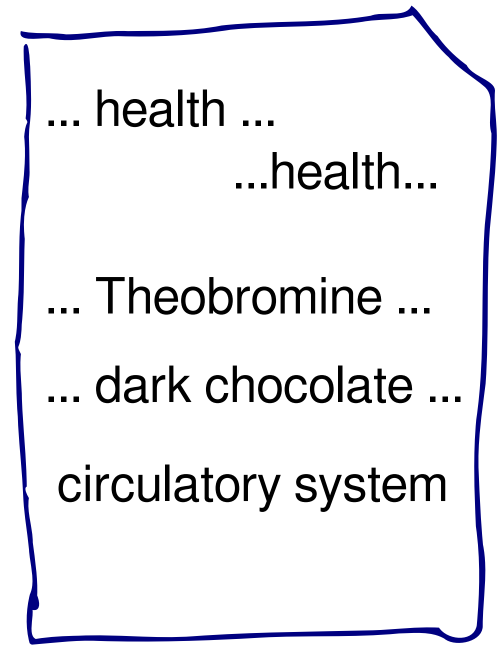


Document relevant?

Matching Entities in Documents by Name

Q: dark chocolate health benefits

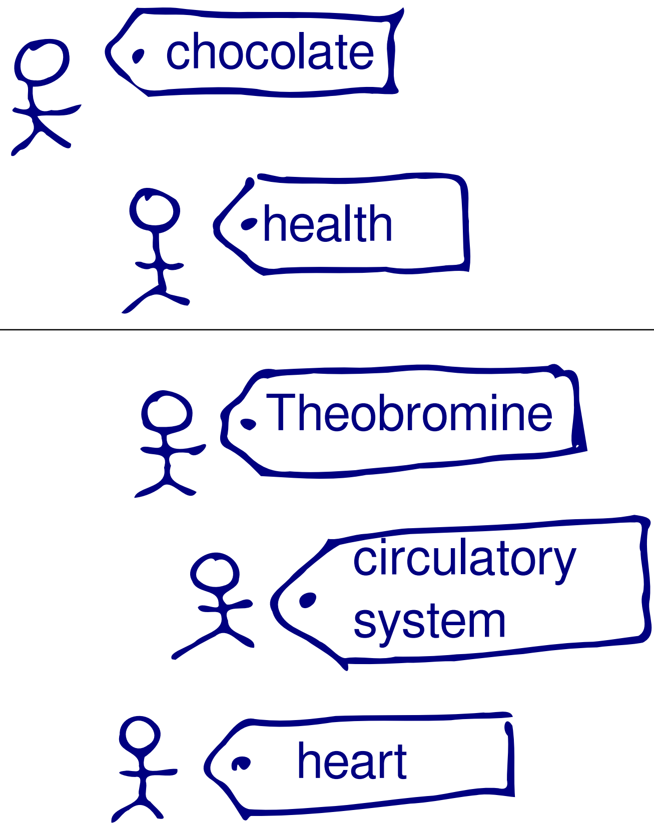


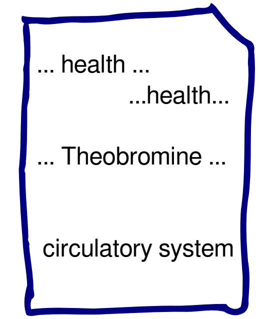


Document relevant?

Matching Entities in Documents by Name

Q: dark chocolate health benefits

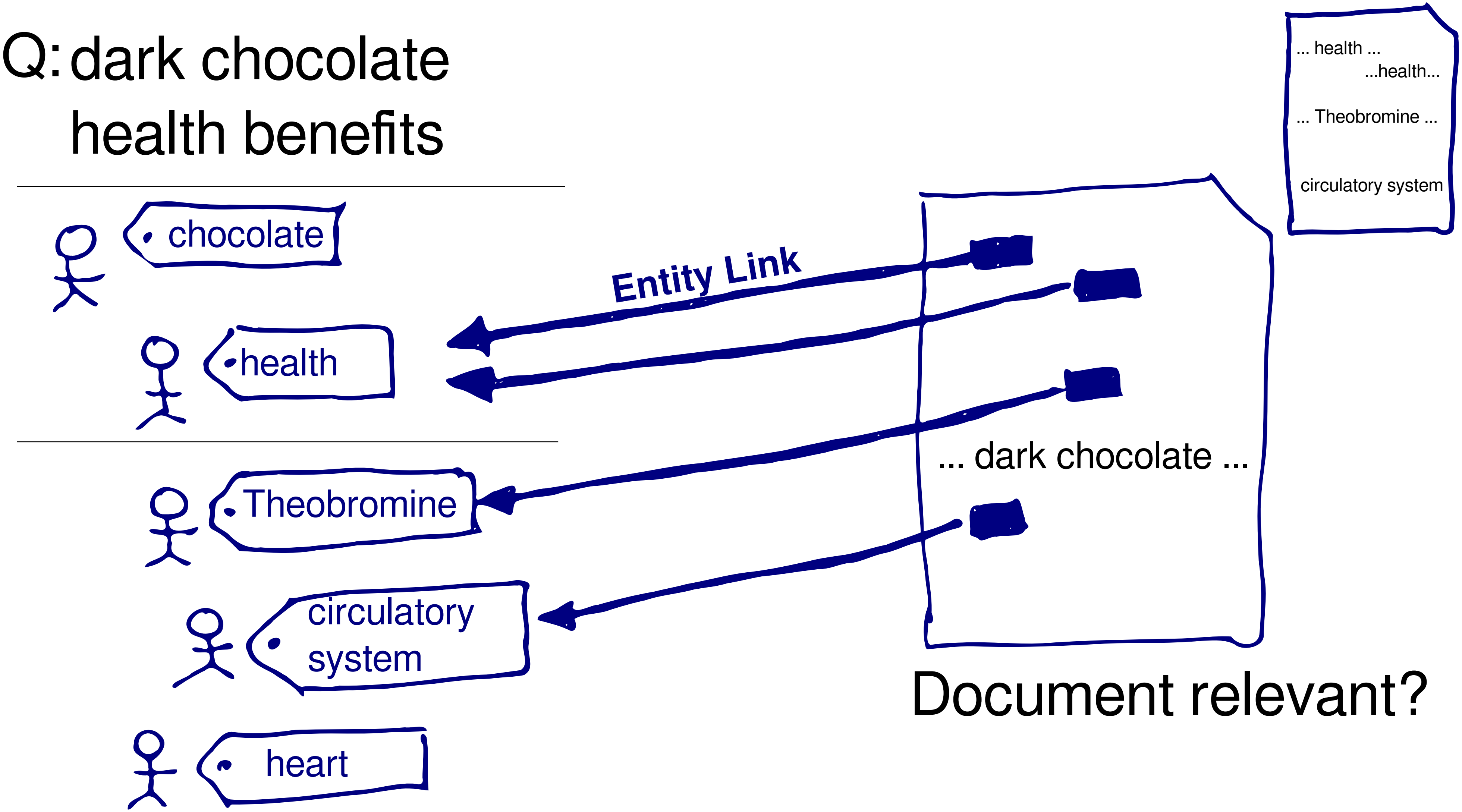




Document relevant?

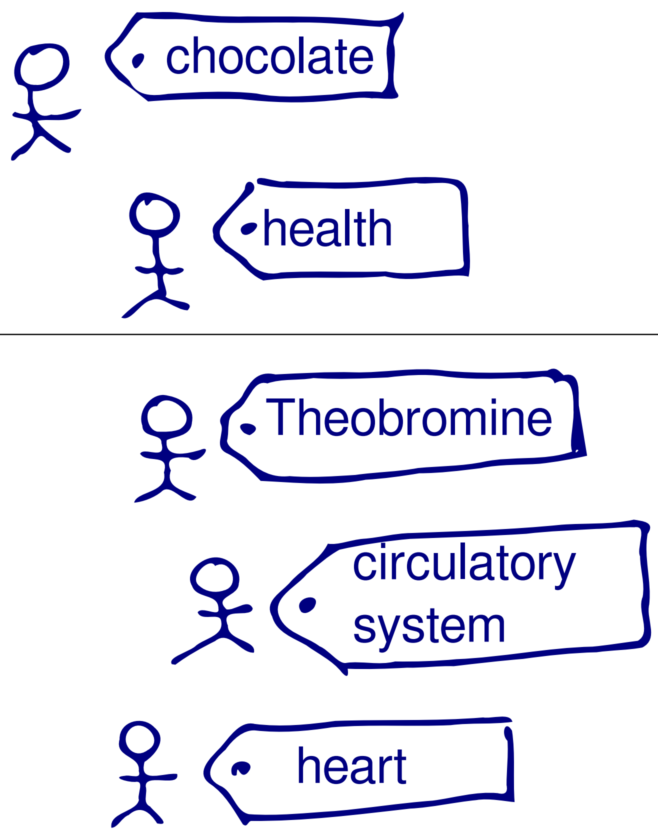
Matching Entities in Documents by Entity Links

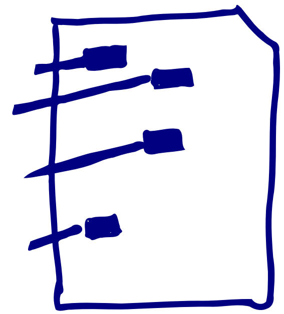
Q: dark chocolate health benefits

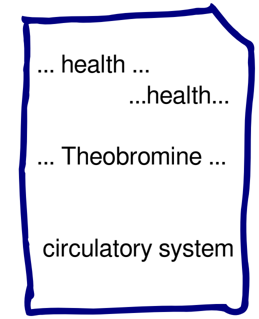


Matching Entities in Documents by Entity Links

Q: dark chocolate health benefits



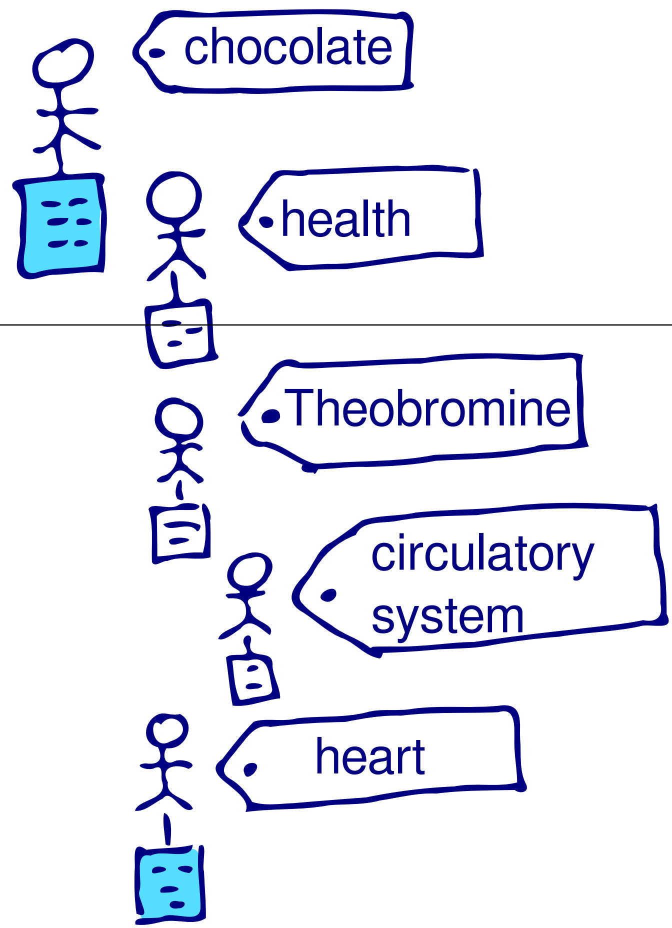


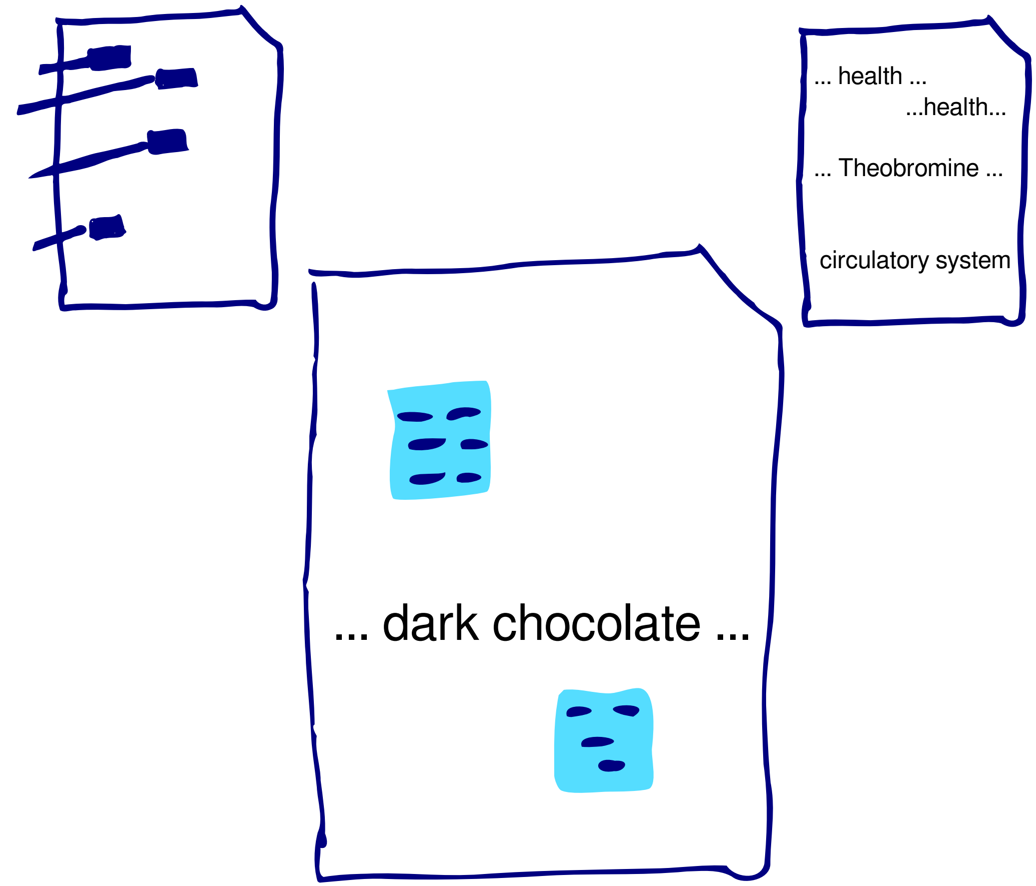


Document relevant?

Matching Entities in Documents by Article Terms

Q: dark chocolate
health benefits

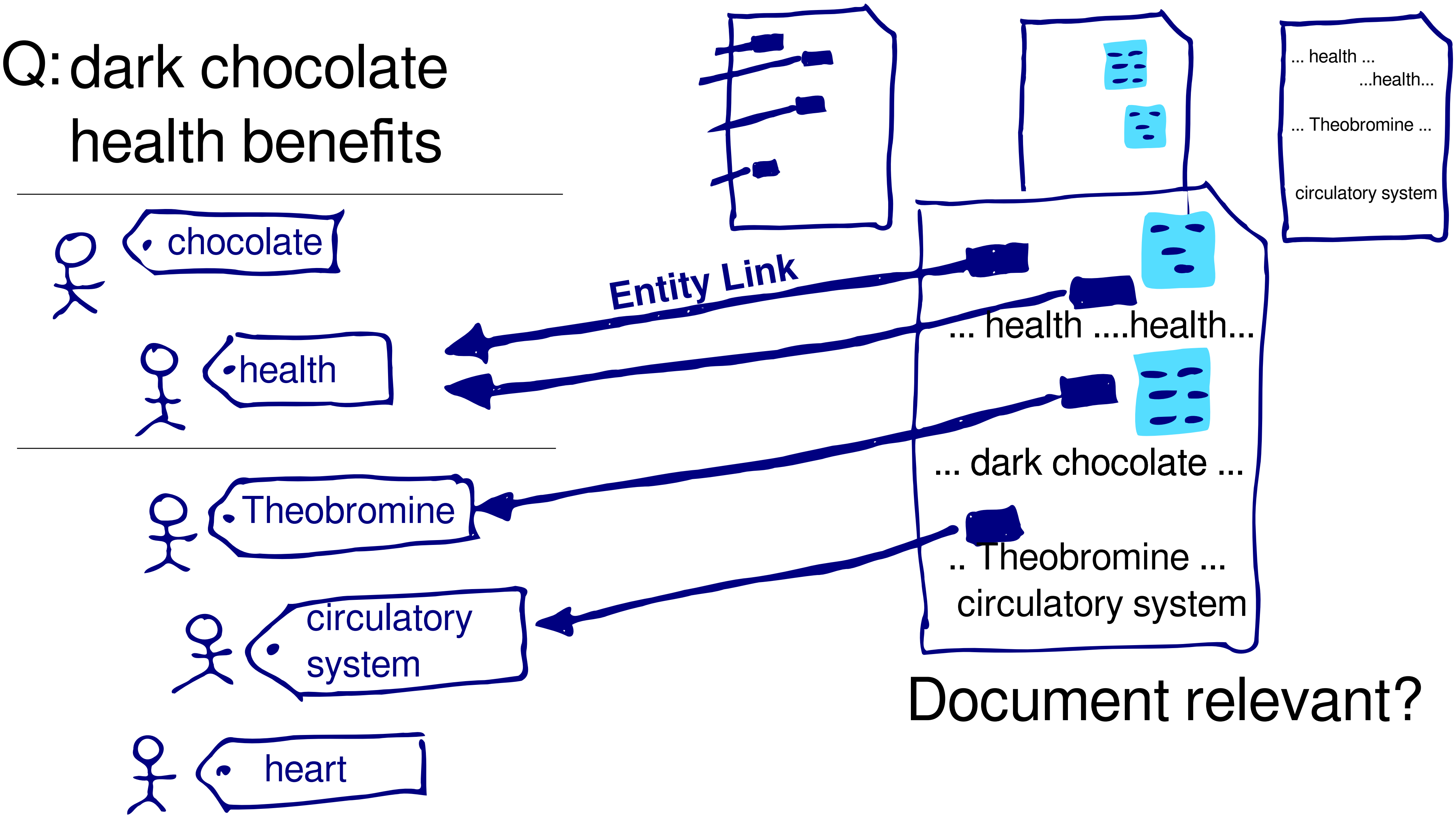




Document relevant?

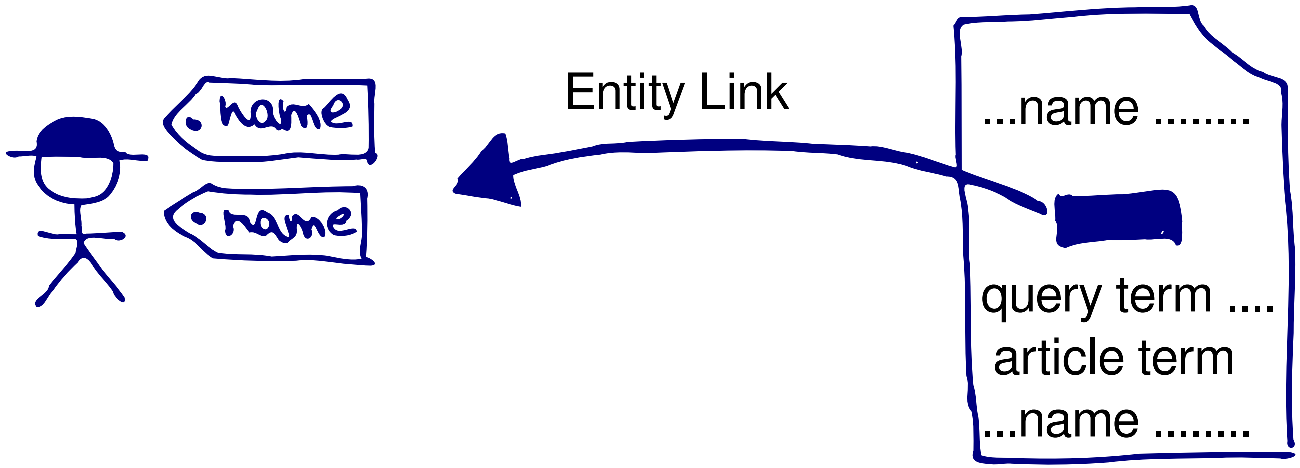
Combine All Names, Links, Terms

Q: dark chocolate health benefits



Document relevant?

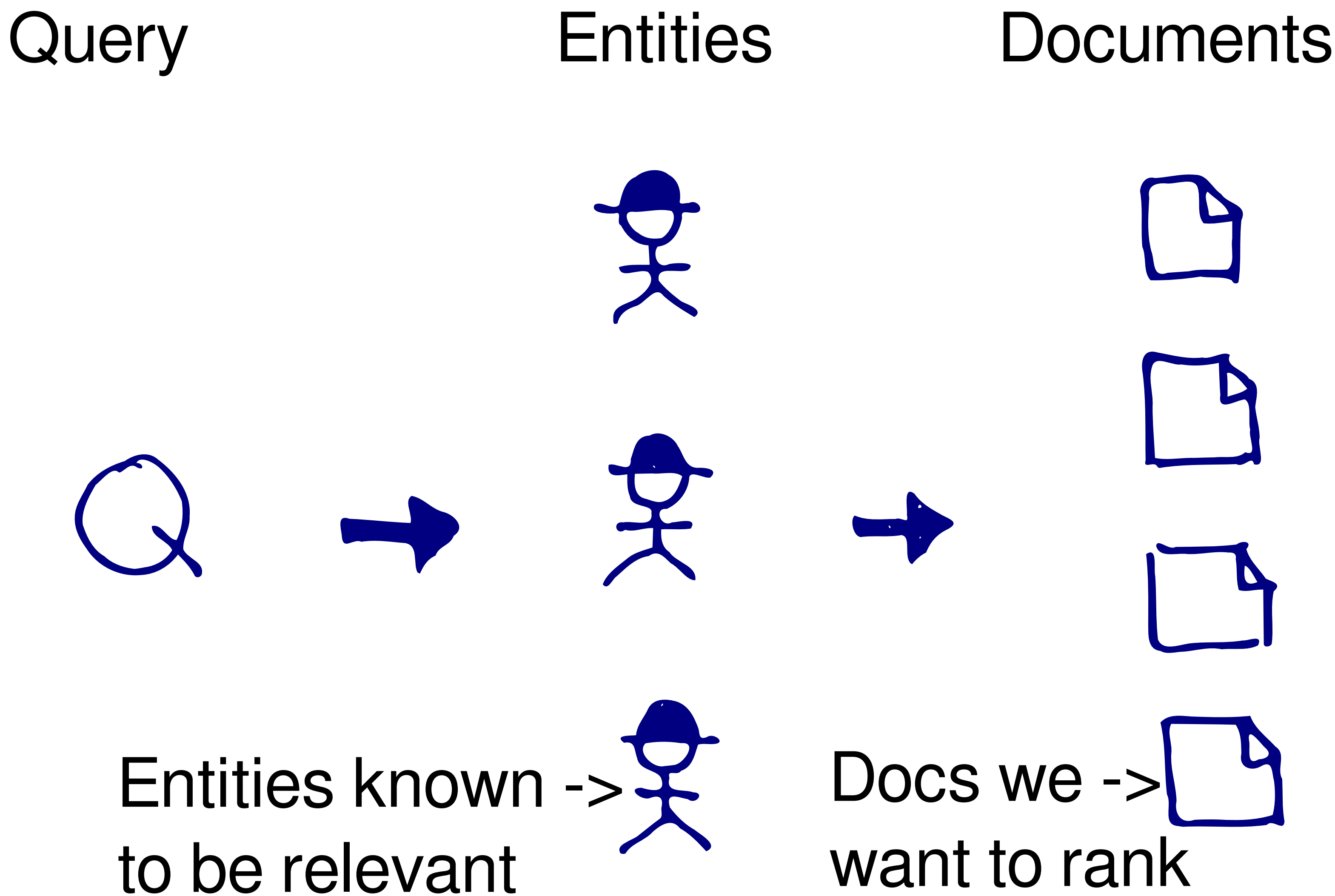
Using Entities as a Vocabulary of Concepts



$$
\begin{aligned}
score(\square) = \quad & \lambda_1 \text{query terms} + \\
& \lambda_2 \text{names} + \\
& \lambda_3 \text{entity links} + \\
& \lambda_4 \text{article terms} + \ldots
\end{aligned}
$$

use your favorite retrieval model here!

Document Retrieval with Entities



Find Relevant Entities

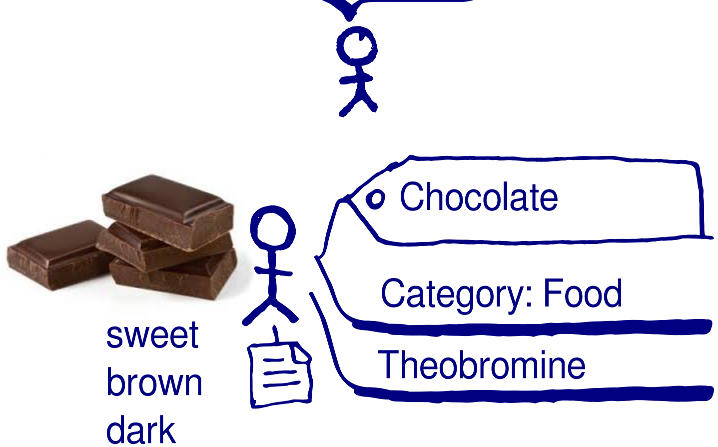
1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

How to Find Relevant Entities?

Query	EU UK relations	dark chocolate health benefits
Query entities	EU UK	chocolate health
Latent entities	Brexit Theresa May	Theobromine circulatory system heart
	Named Entities	**Concepts**

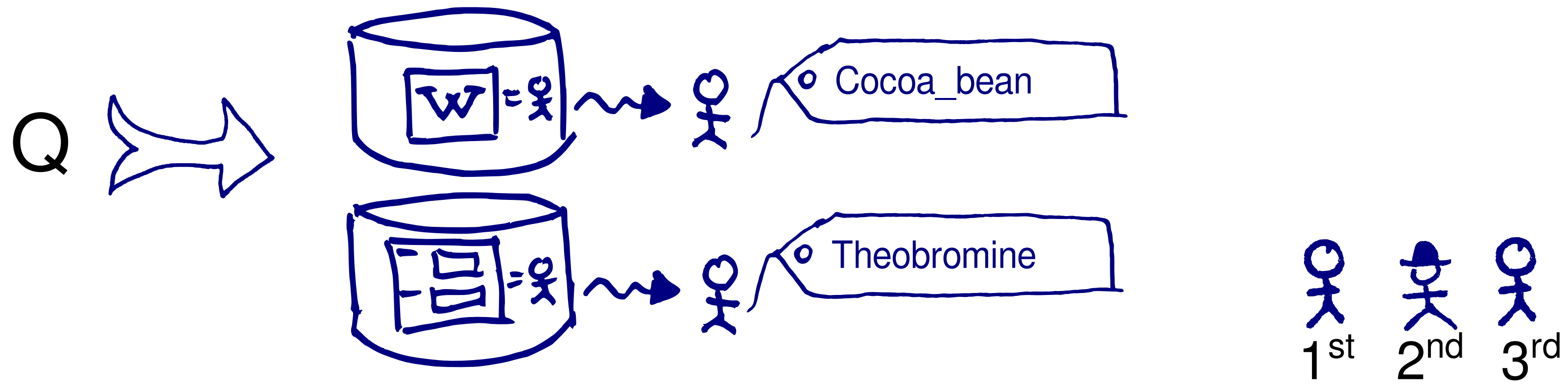
Query Entities through Entity Linking

Query: dark chocolate health benefits



Latent Entities through Retrieval (e.g., Part 3)

Retrieve entities from knowledge base
to obtain ranking of entities E (with score)

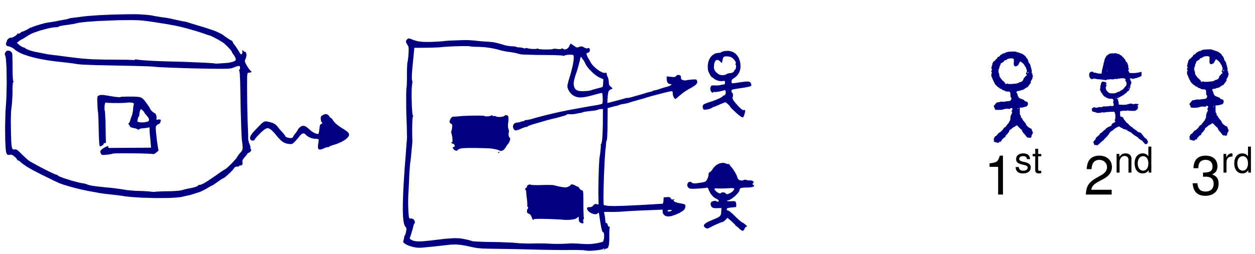


Latent Entities through Pseudo-Relev. Feedback

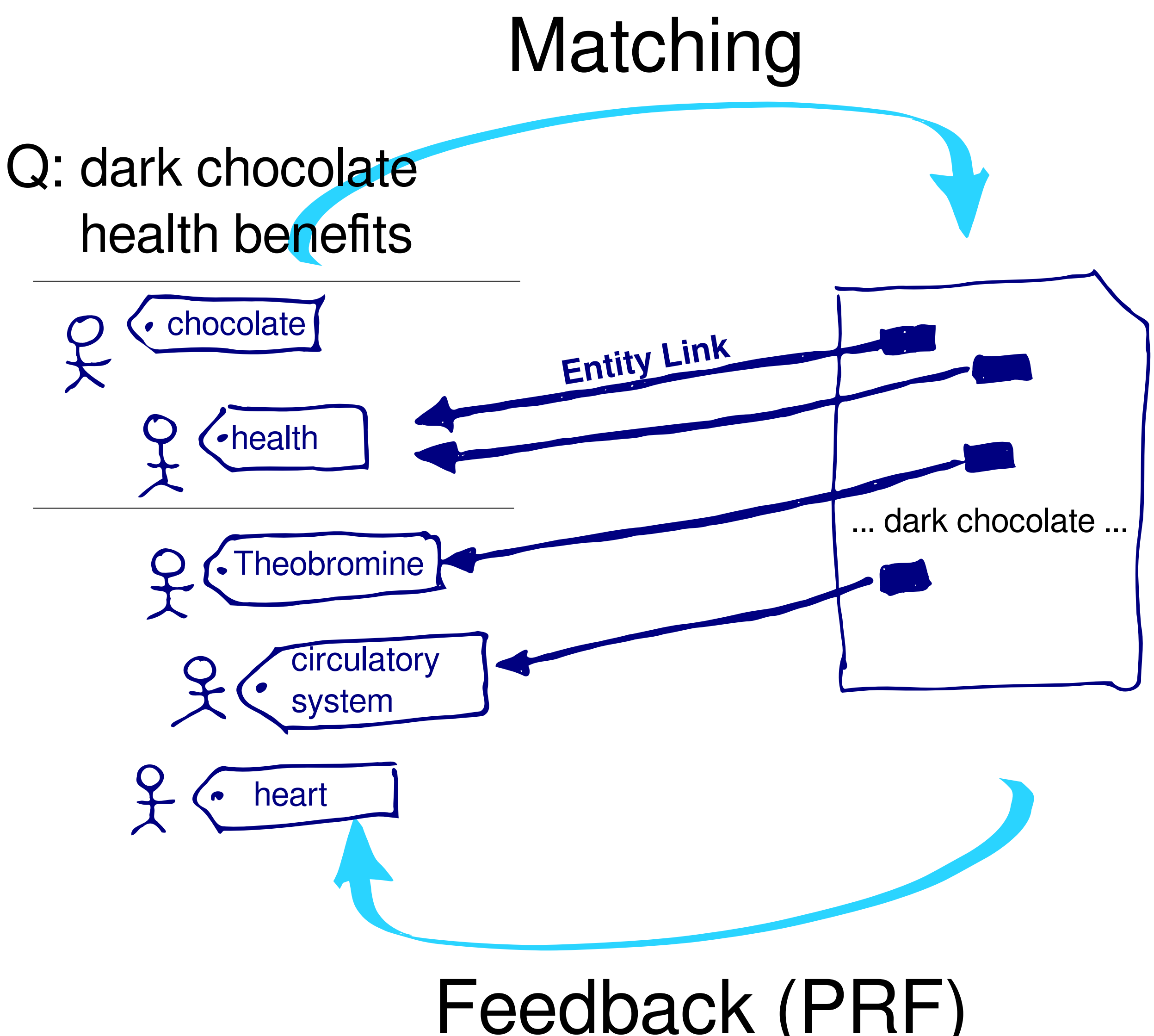
1. Retrieve preliminary documents
2. Entity link documents
3. Derive distribution over  (bag of entities)

(see Relevance Model / RM3)

[Dalton et al 14, Liu & Fang 15]



PRF is Inverse of Matching Entity Links

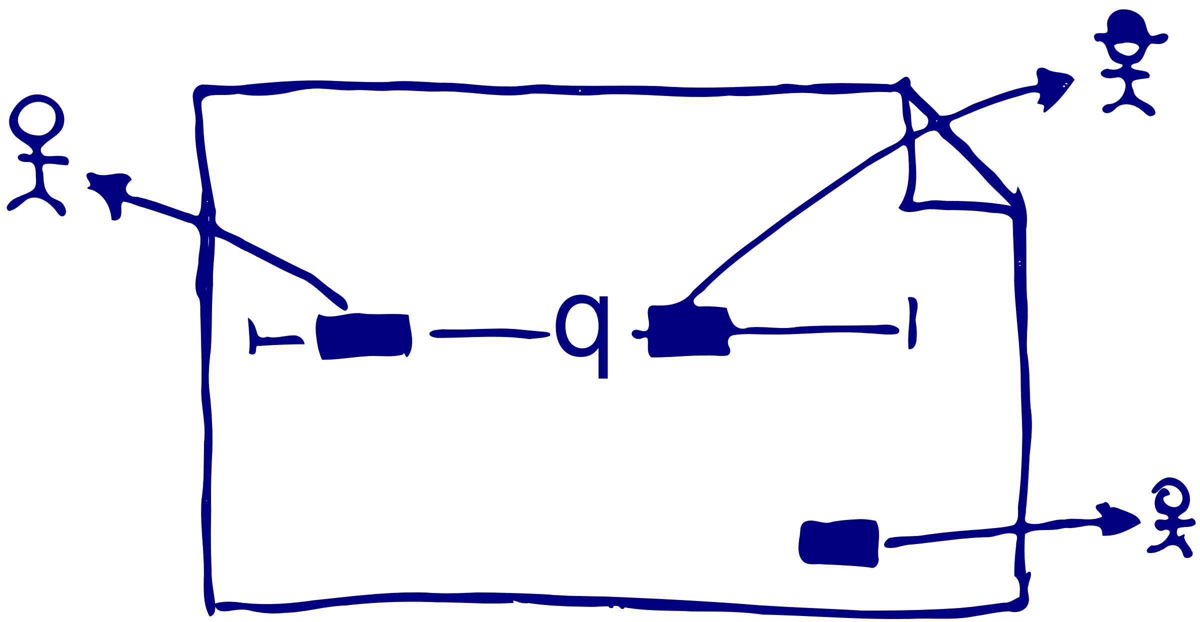


Latent Entities through Proximity to Query Words

Using distance between entity mentions and query words **q** as a measure for relevance.

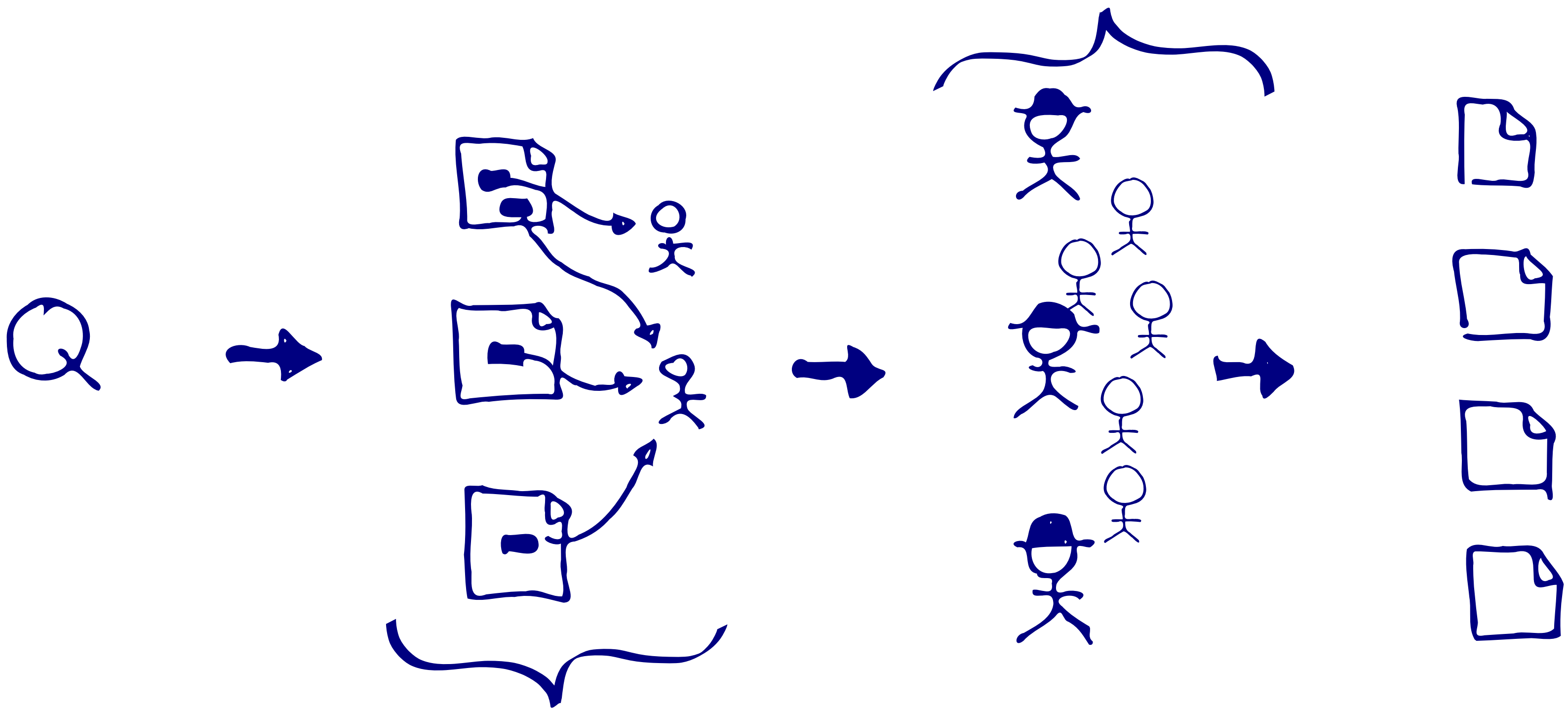
[Petkova & Croft, 07]

also see enity profiles [Liu & Fang, 15]
entity context model [Dalton et al, 14]



Entity Expansion for Document Retrieval

Query entities + Object retrieval (Part 3)

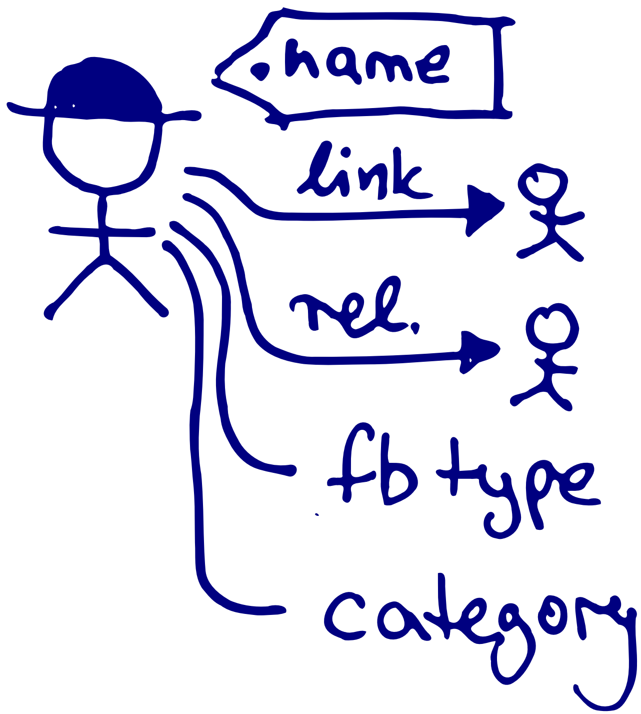


Pseudo-relevance feedback (RM3)

Document = bag of entity links (instead of terms)

Using More from the Knowledge Graph

So far we used names and entity links.
But KGs have so much more information!



Names

Links and relations

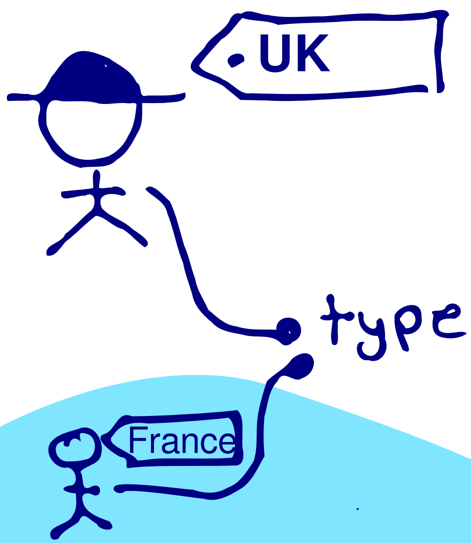
Different taxonomic type systems

How can we make use of it?

Using Types and Categories

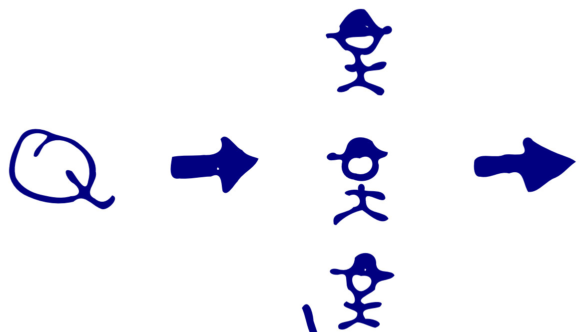
1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

Utilizing Entity Types



inferred as relevant
because of same type

Entity Types Inferred through Entity Links



Which types are relevant ?

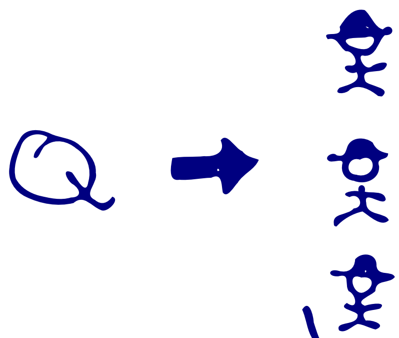
How to match types to documents ?

a) same-type entities
[Kaptein et al 10]

majority types among entities

prefer docs with entities of this type

Entity Types Inferred through Entity Links



Which types are relevant ?

How to match types to documents ?

a) same-type entities
[Kaptein et al 10]

majority types among entities

prefer docs with entities of this type

Method	MAP on INEX
Full Text	0.03
Link	0.09
Type+Link	0.13

Entity Types through Text Classification

Q →

Which types are relevant ? →

How to match types to documents ?

b) term classifier [Xiong & Callan 15]

classify query terms with naive Bayes

classify documents with naive Bayes

Open Issues Regarding Types

- Often types are very broad
- Often entities of many different types are relevant
- Often some entities of a type are relevant,
  others are not...

Quality issues of type ontologies

- Wikipedia categories are very noisy

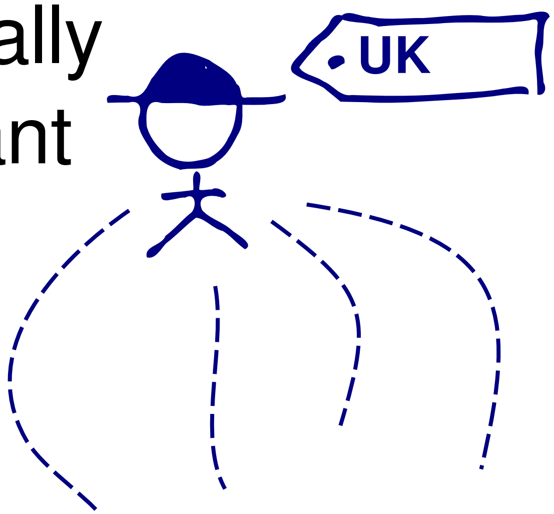
Graph Expansion

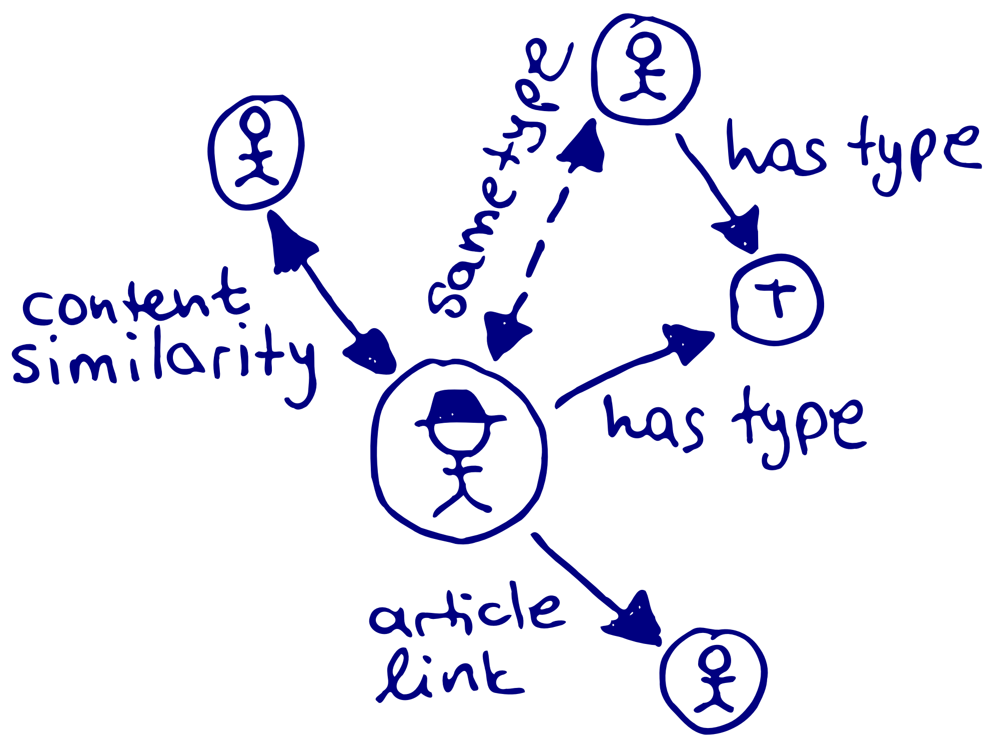
1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

More Entities Found near Relevant Entities

Query: EU UK Relations

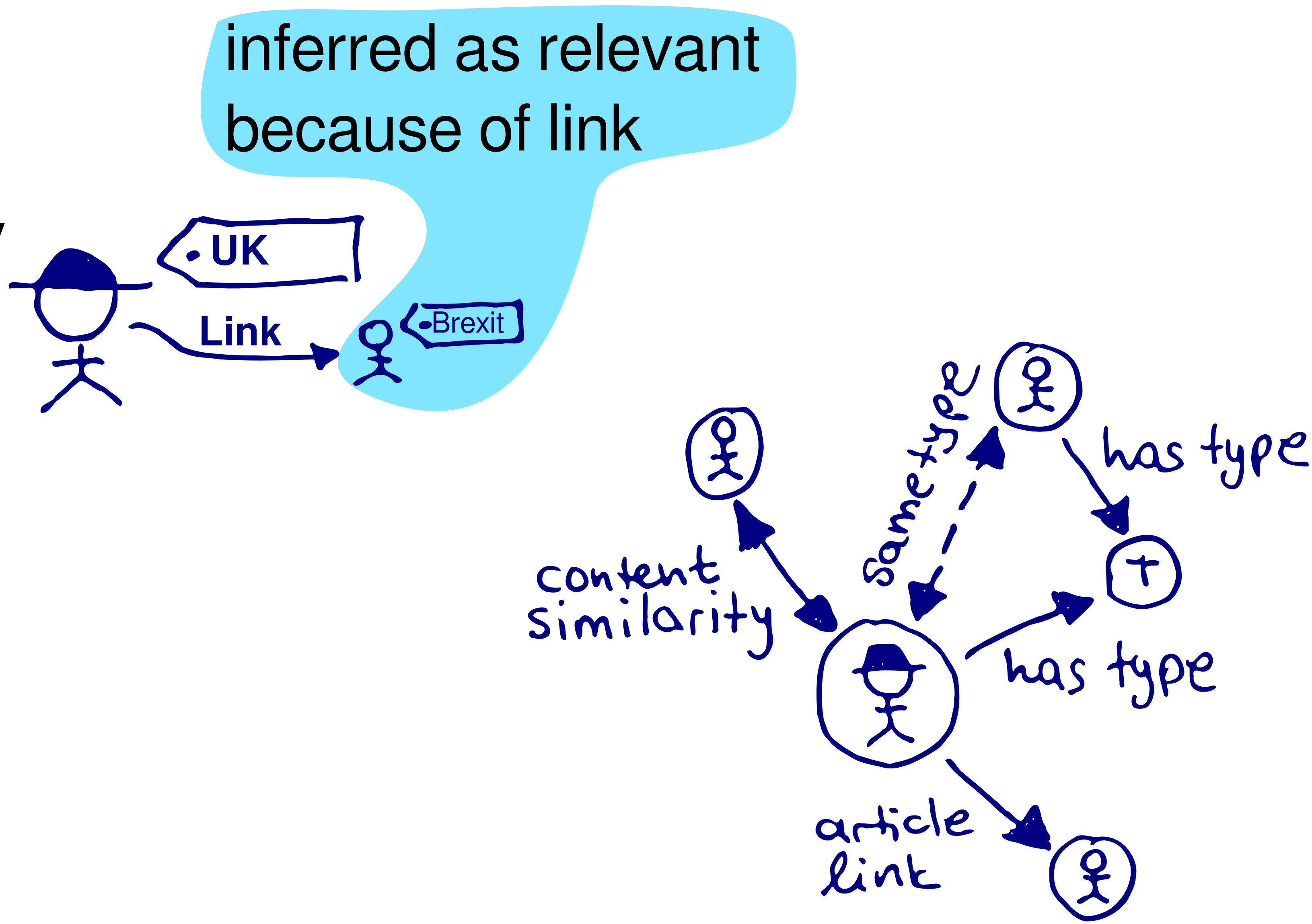
originally relevant



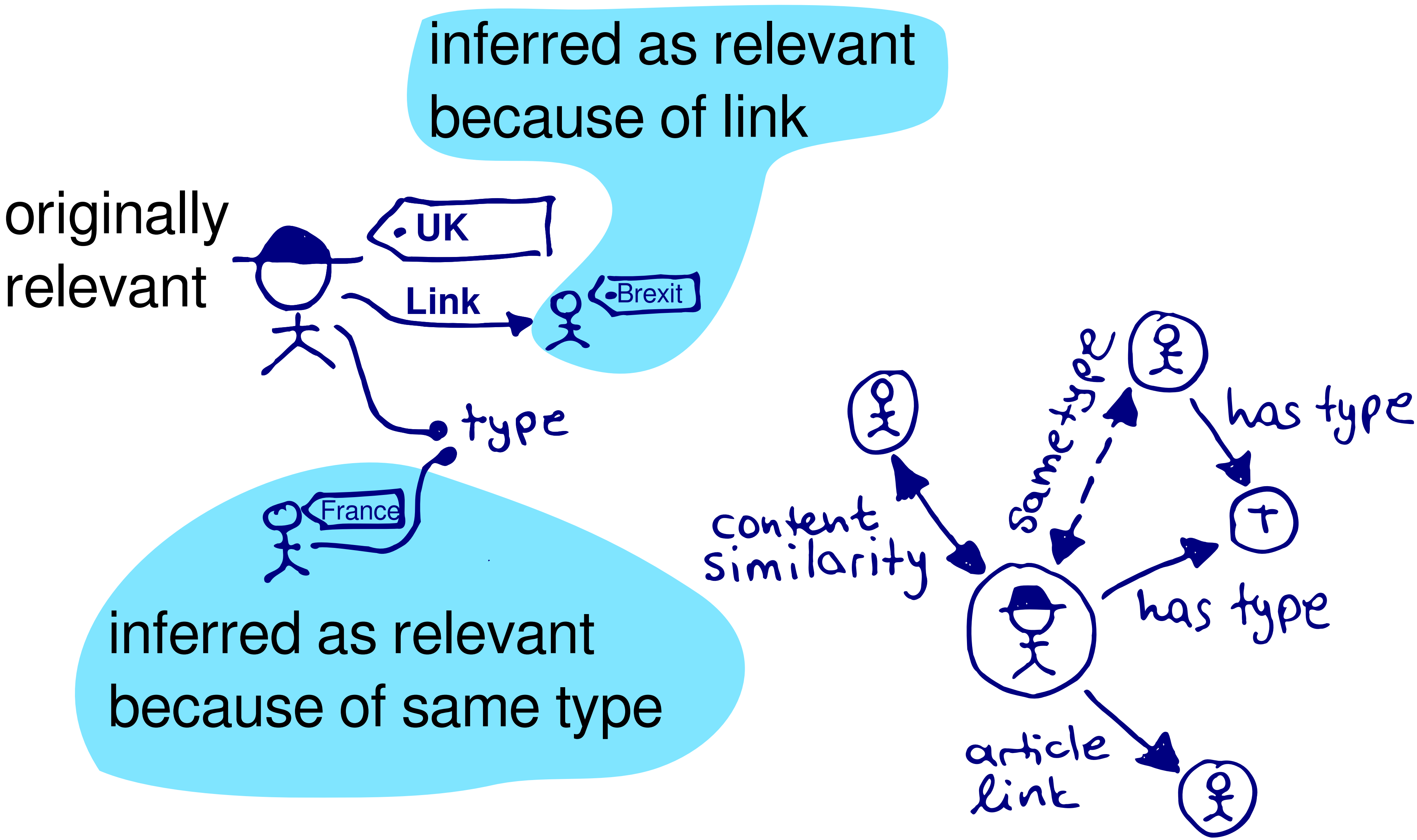


Using Relations and Types with Entity Links

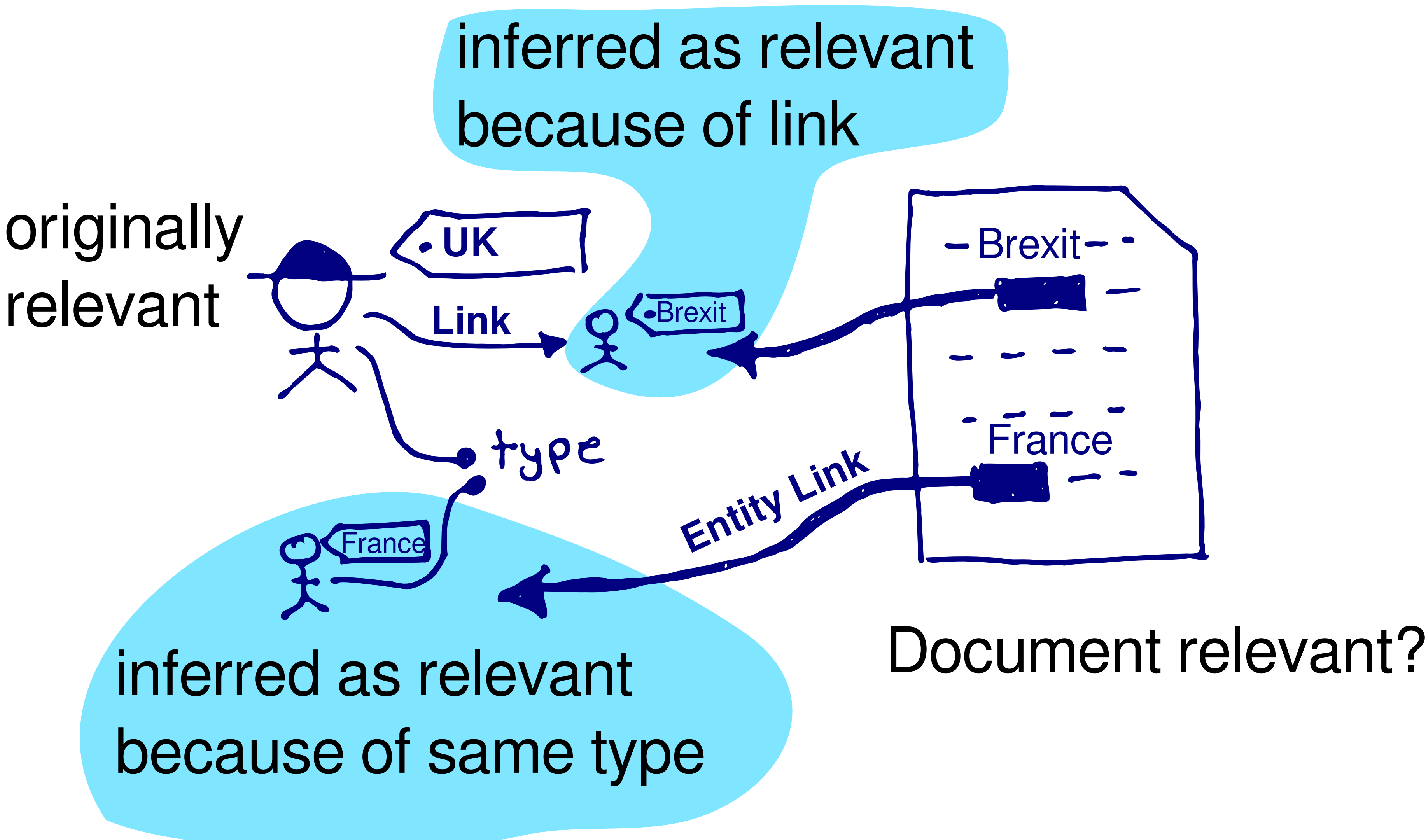
originally
relevant



Using Relations and Types with Entity Links

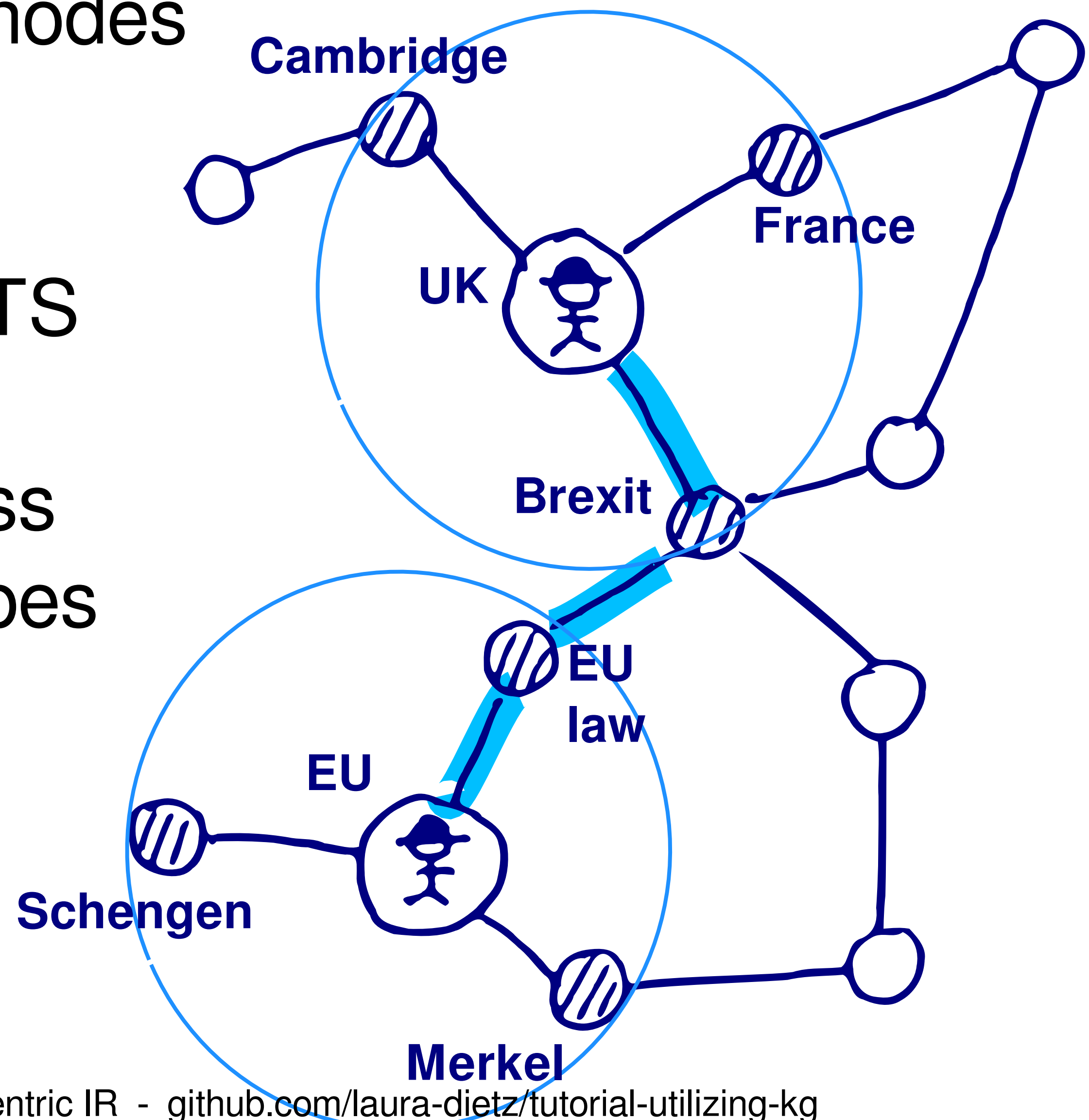


Using Relations and Types with Entity Links



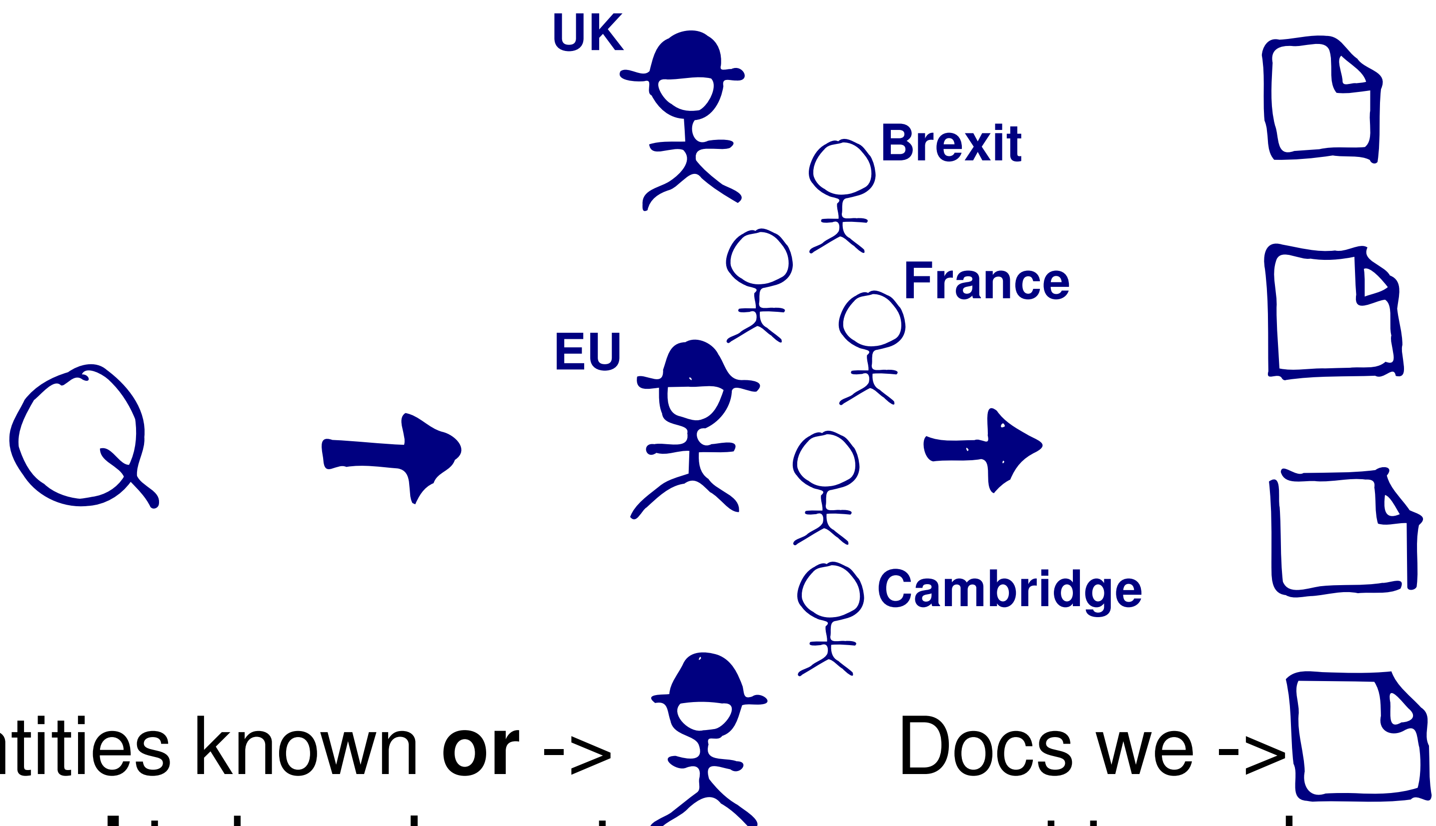
Graph Expansion

Using seed entity nodes
- 2-hop expansion
- Graph walks:
  - PageRank / HITS
- Shortest Paths
- Entity Relatedness
- Different edge types
- Edge weighting
- Graph Clustering



Document Retrieval with (More) Entities

Query | Entities | Documents



Entities known **or** -> **assumed** to be relevant

Docs we -> want to rank

Successes of using the Link Structure

Kaptein et al 2010

Kotov & Zhai 2012

Boston et al 2014

Weight entities by:

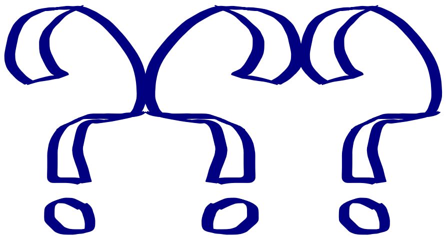
M: How well **E**s article content matches the query

MR: How often **E** is linked by others (PageRank)

Method	F1 on TREC QA
M	76.92
M+d*MR	79.47

(d=0.0001)

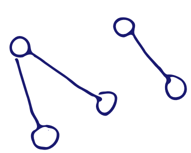
In recent years, links seem to not work any more... ?!?



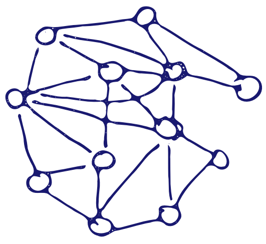
Link Structure "Stopped Working"

Wikipedia started with the "most popular" facts
then it grew in number of nodes and
number of connections, aiming for better coverage.

Wikipedia in 2013



Wikipedia in 2018

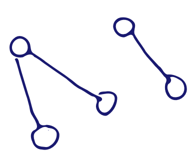


Hub nodes: New York City, California, United States

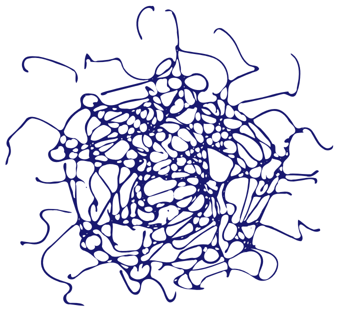
Link Structure "Stopped Working"

Wikipedia started with the "most popular" facts
then it grew in number of nodes and
number of connections, aiming for better coverage.

Wikipedia in 2013



Wikipedia in 2018

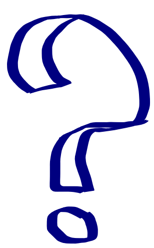


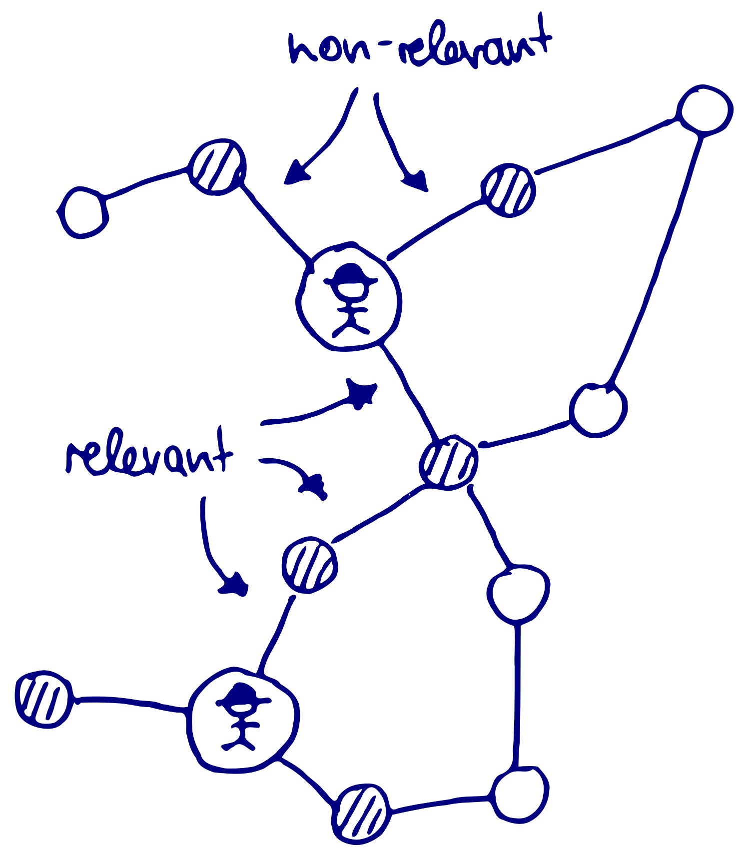
Hub nodes: New York City, California, United States

Big Question

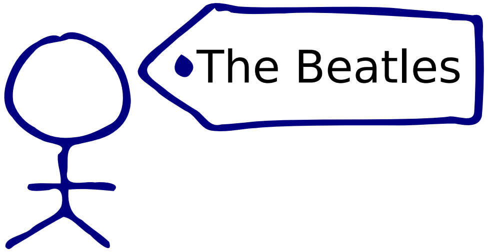
How to infer which other connected entities / nodes are relevant for the information need Q?

...and therefore safe for expansion?



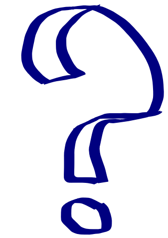


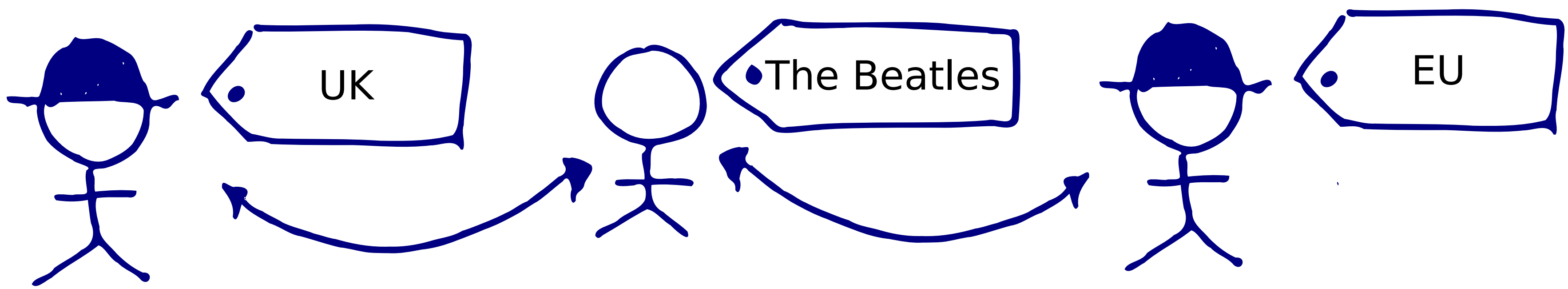
Open Issue: Predict Relevance of Edges



Open Issue: Predict Relevance of Edges

An edge can be relevant for one query
and non-relevant for another query
Can't be distinguished through graph structure.

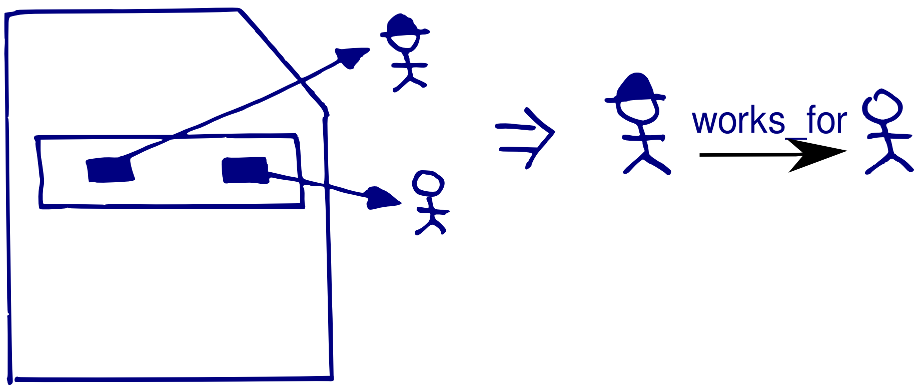




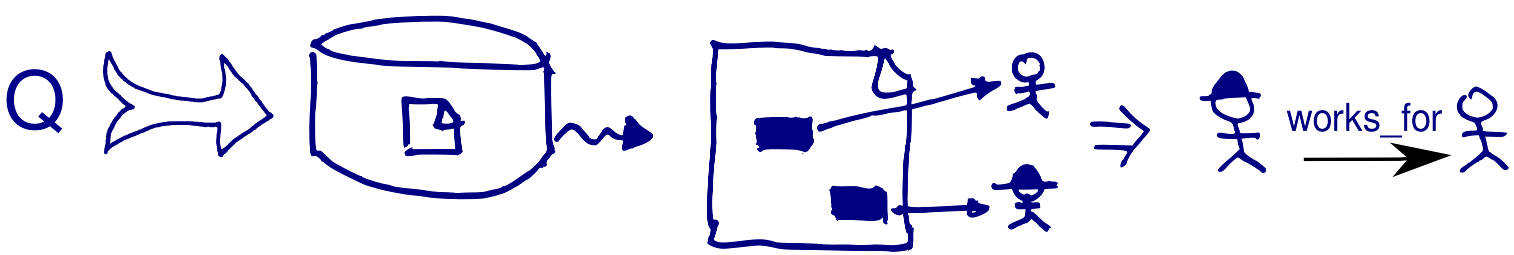
Edges are relevant for query: EU UK Bands
Edges are **not** relevant for query: EU UK Relations

Using Relation Extraction

Relation Extraction:
[Roth et al 14]
(best at TAC KBP 13)



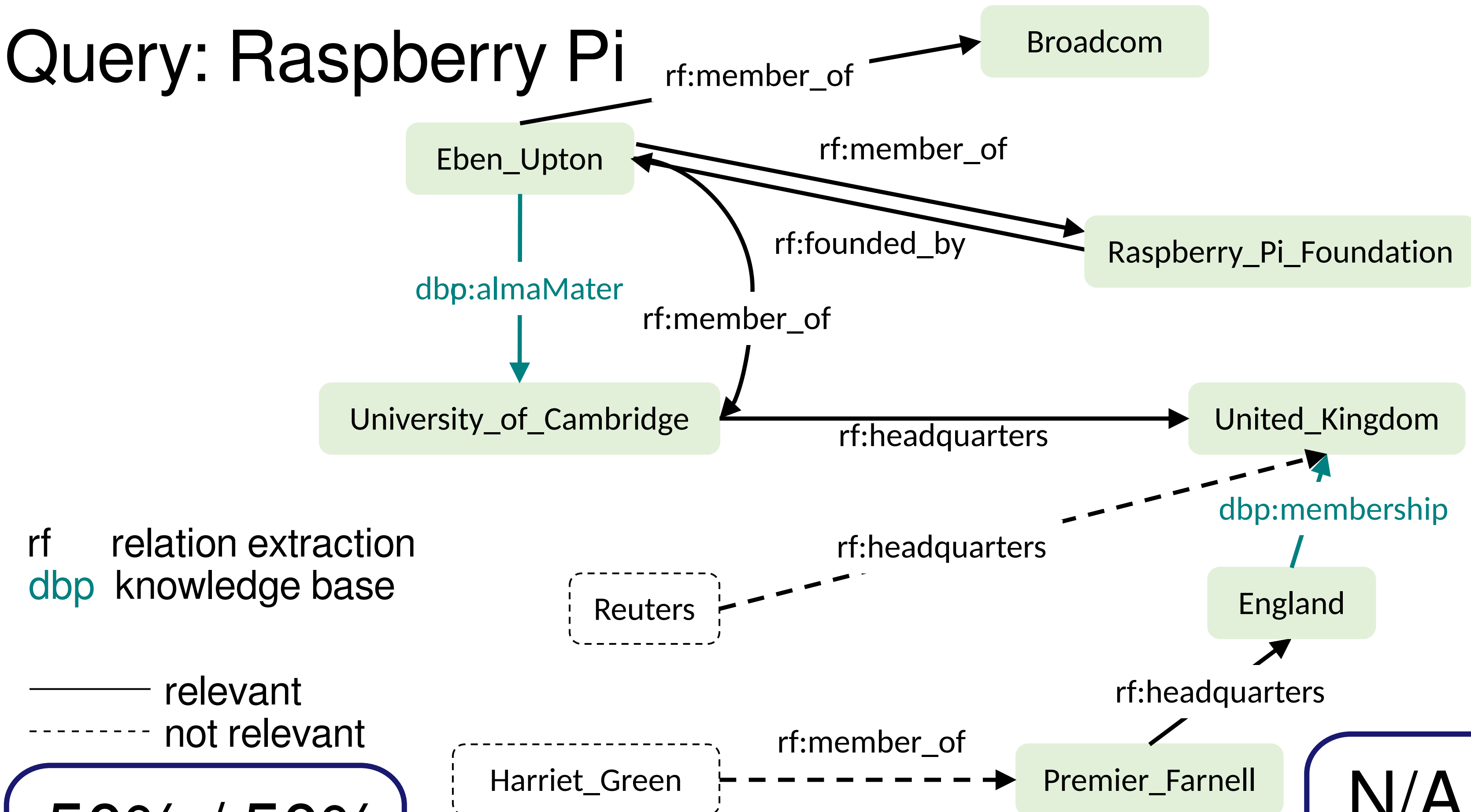
Research question:
relevant documents + extraction = relevant relations?
[Schuhmacher, Roth, Ponzetto, Dietz 16]



Relations of Relevant Documents [Schuhmacher et al, 16]

Goal: Relations need to be relevant and correct

Query: Raspberry Pi

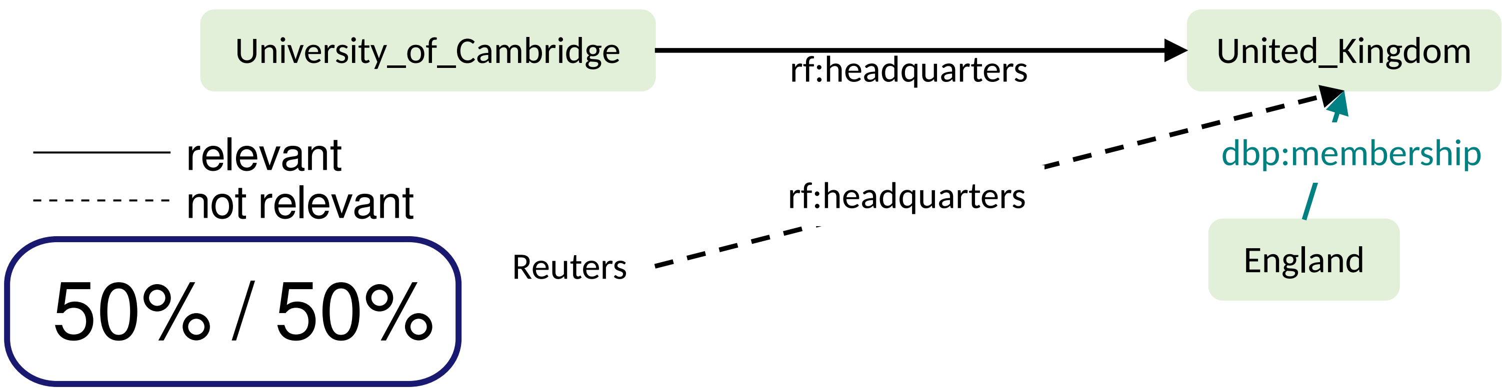


50% / 50%

N/A 60% queries

More Big Questions

How to deal with high number of non-relevant relations in relevant documents?



How to utilize relation types, when the query does not explicitly mention them?

Combination of Multiple Sources

1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

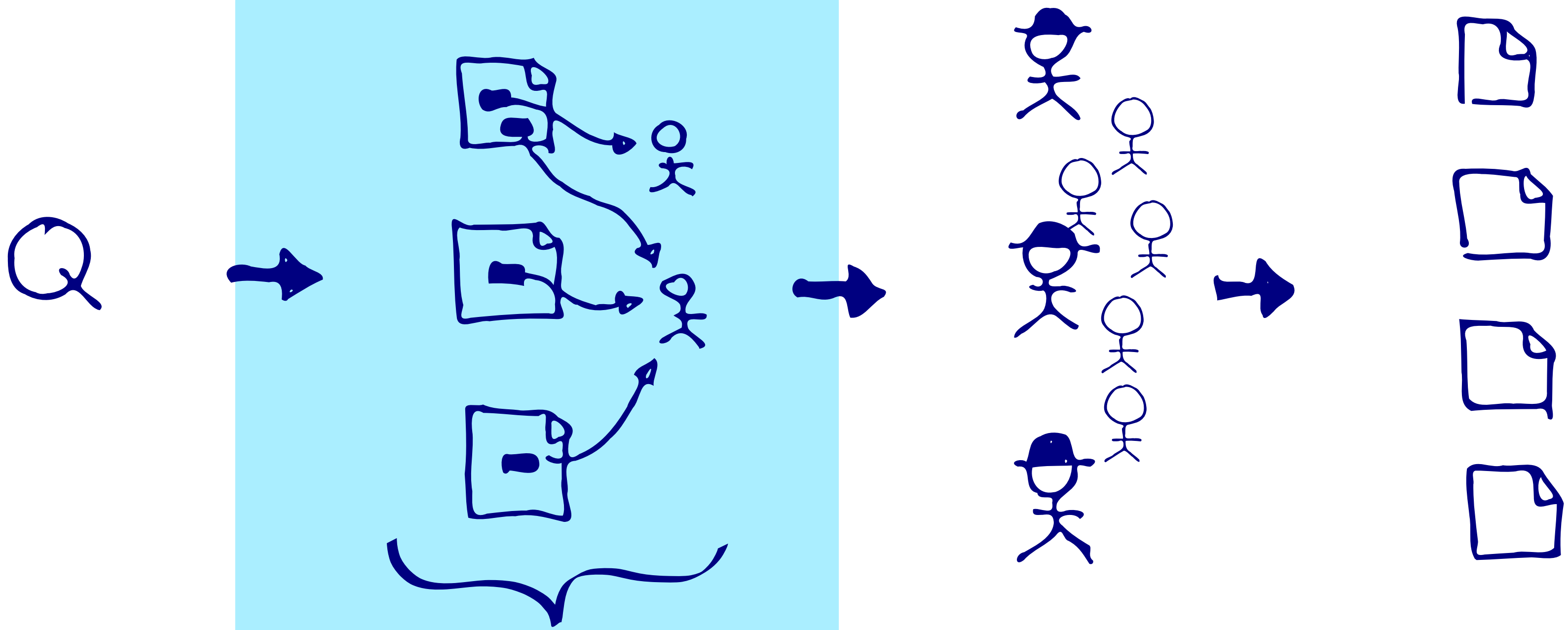
Complementary Sources

Typical approaches:

1) Use **complementary sources**:
   graph, article text, relevance feedback, type info

2) Use **machine learning**:
   Train weights for sources on test collection

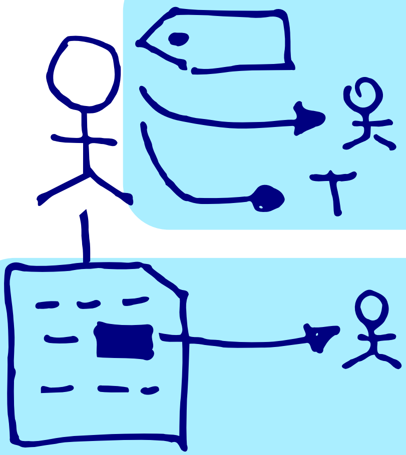
Source: Relevance Feedback with Entity Links



Pseudo-Relev.
Feedback (RM3)

Document = bag of Entity Links
Proximity of query and Entity Links
[Petkova 2007, Dalton et al 14, Liu & Fang 15]

Source: Object AND Article Content Retrieval



Entities as attribute-structured objects:
Object retrieval (see Part 3 & [Hasibi et al 16])

Entities as text:
Each article represents an Entity
Retrieve articles with keyword query Q
=> ranking of entities

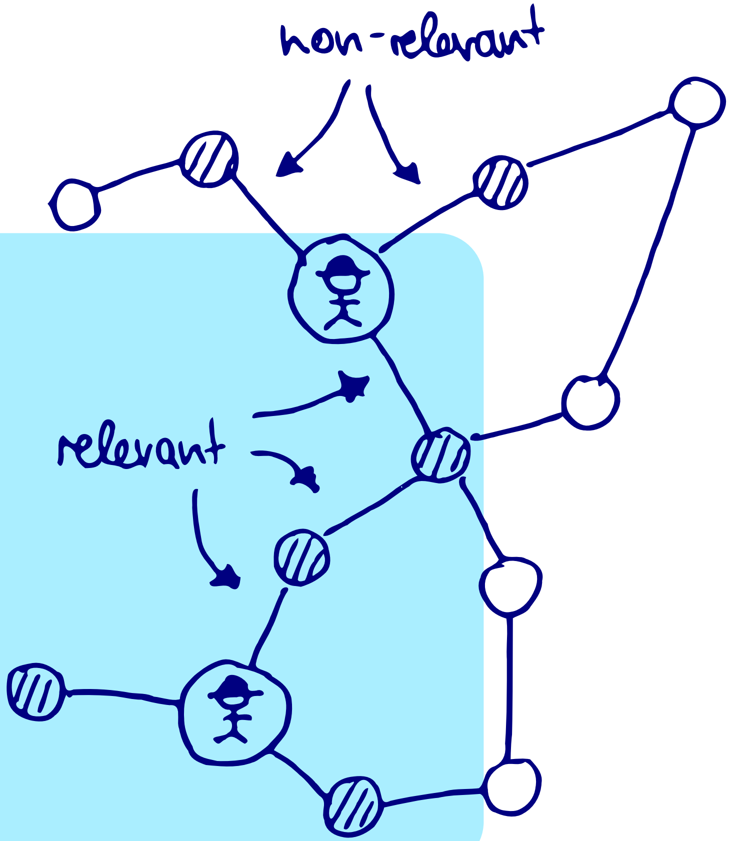
[Xiong & Callan 15, Dalton et al 14]

Source: Graph Structure and Walks

Graph Walks

[Boston et al 2014,

Kotov & Zhai 2012]



Machine Learning

1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

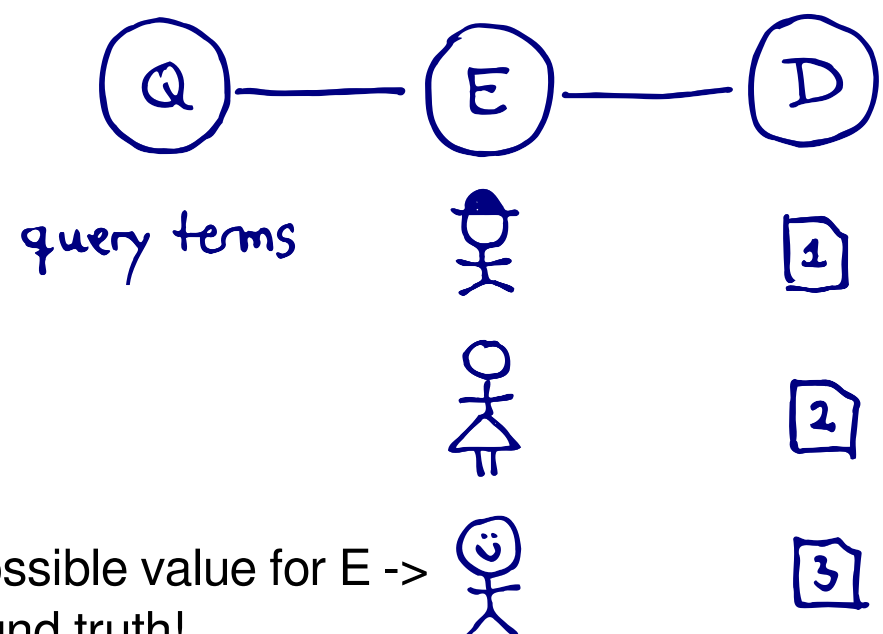
Machine Learning / Probabilistic Models

Three approaches based on similar ideas:

- Dalton: Entity Query Feature Expansion
- Xiong: EsdRank
- Liu: Latent Entity Space

Probabilistic model with random variables Q,E,D.

An edge represents a measure of compatability or similarity.

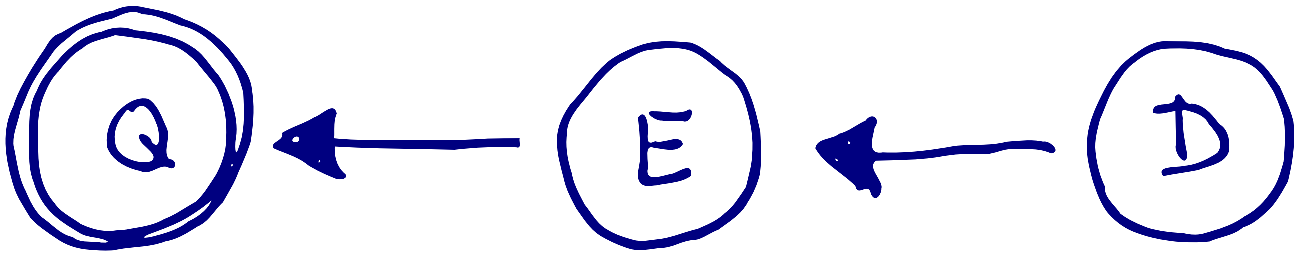


One possible value for E -> no ground truth!

<- One possible value for D ground truth available (TREC)

Latent Entity Space

[Liu & Fang 15]



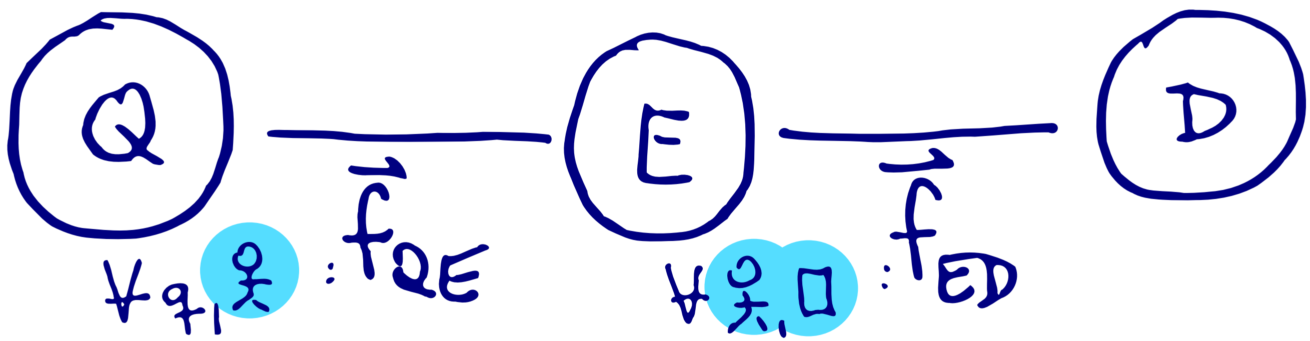
$$p(q|D=d, R=1) = \sum_{e \in \mathcal{E}} p(q|e) \cdot p(e|d)$$

similarity of LM(q) and LM(e)

similarity of LM(e) and LM(d)

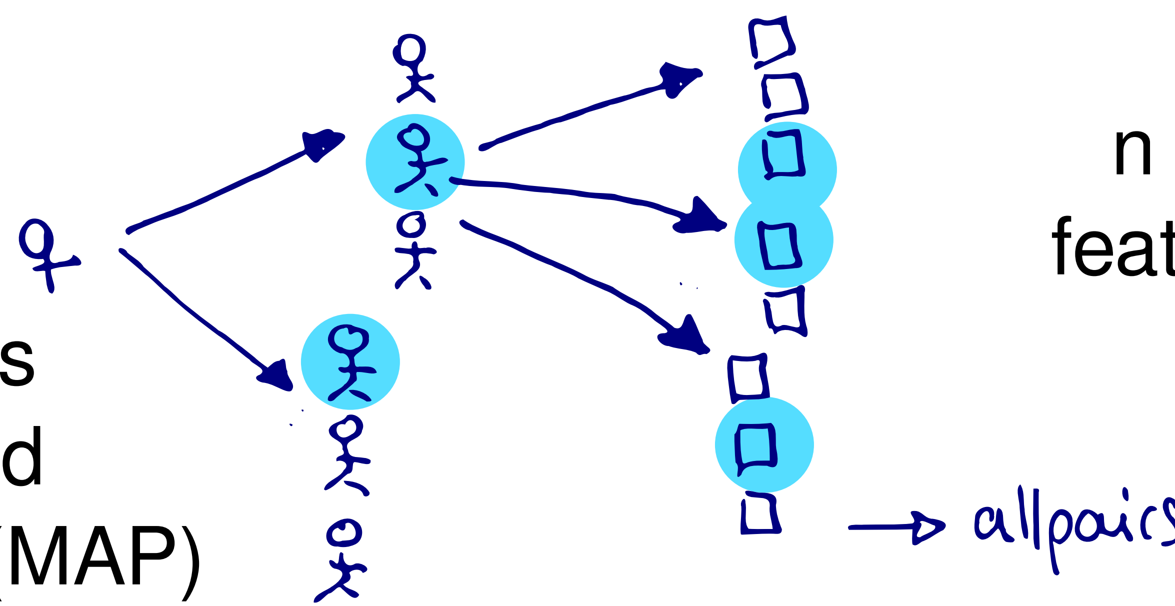
Wide range of experiments on which similarity measure / data source combination works best.

Entity Query Feature Expansion [Dalton et al, 14]

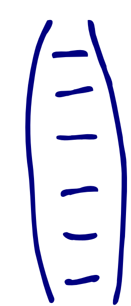


n different ways to compute p(q|e)

m different ways to compute p(e|d)



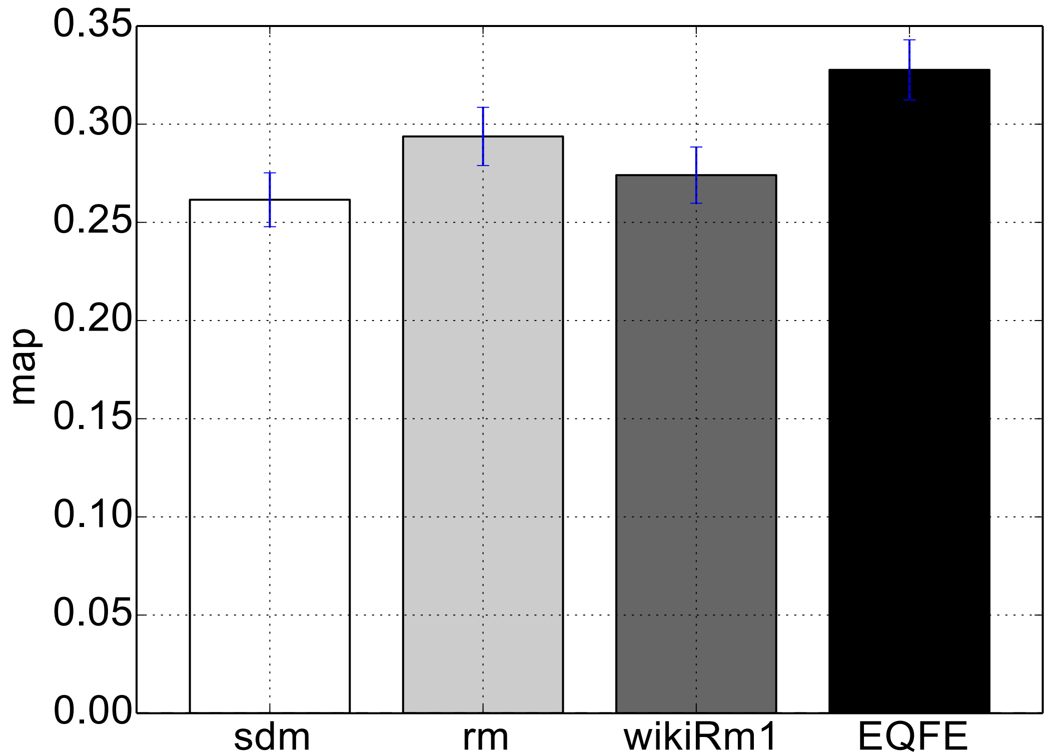
n x m features!



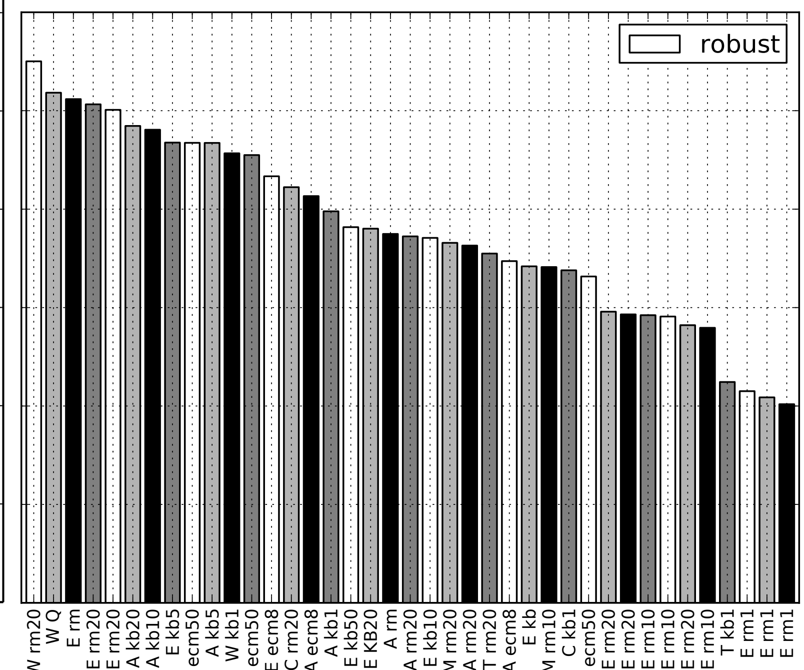
Combine features then use standard learning to rank (MAP)

Entity Query Feature Expansion [Dalton et al, 14]

Results on Robust04 ad hoc document retrieval.

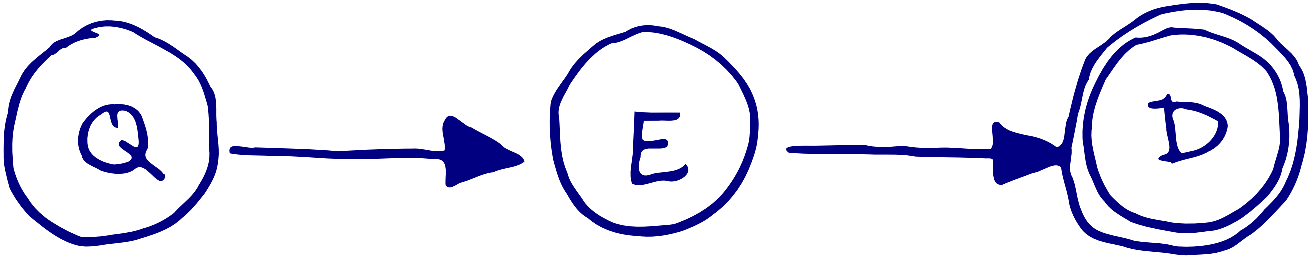


combined



features

EsdRank [Xiong & Callan 15]



$$p(d_i|q) = \sum_{e \in \mathcal{E}} \underbrace{p(d_i|e)}_{\frac{1}{Z_1}\exp\left\langle \vec{w}_1, \vec{f}_{D,E} \right\rangle} \cdot \underbrace{p(e|q)}_{\frac{1}{Z_2}\exp\left\langle \vec{w}_2, \vec{f}_{E,Q} \right\rangle}$$

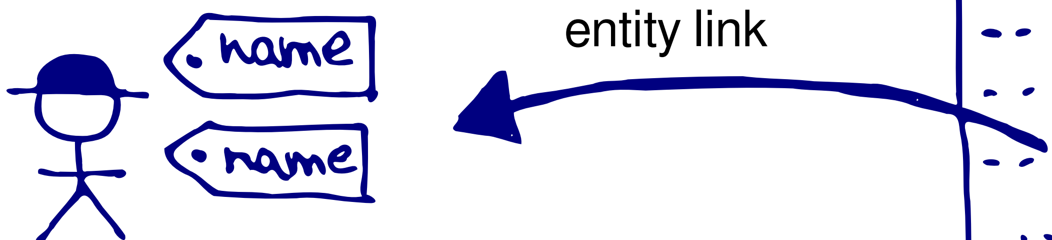
Discriminative probabilistic model based on Generalized linear models + EM Algorithm for learning weights w1, w2.

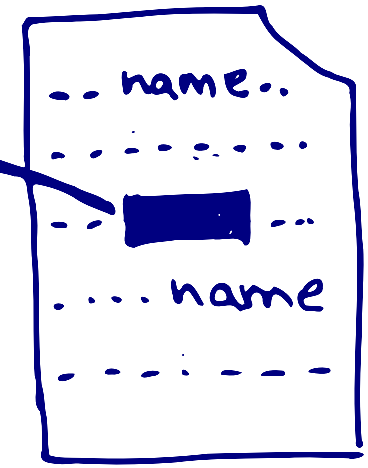
Only n+m features! But needs custom learning code.

[Metzler & Croft 07]

Relation to Query / Latent Concept Expansion

Various vocabularies, but all represented by sets





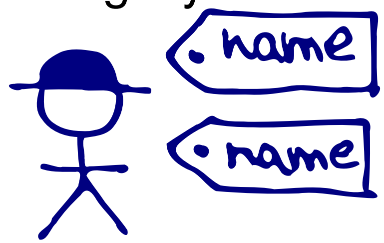
$$
\begin{aligned}
score(\square) = \quad & \lambda_1 \text{query terms} + \\
& \lambda_2 \text{names} + \\
& \lambda_3 \text{entity links} + \\
& \lambda_4 \text{article terms} + \ldots
\end{aligned}
$$

Query Expansion with Uncertainties

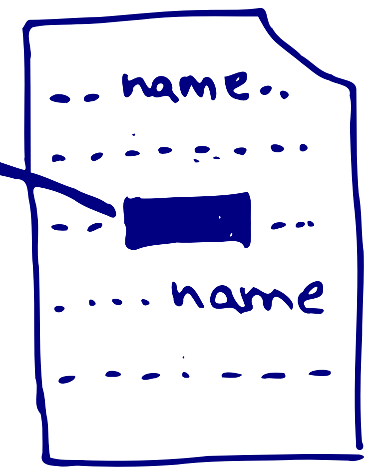
Taking uncertainty and confidences into account.

[Raviv et al 16, Liu & Fang 15]

Ambiguity of names



uncertainty of links



uncertainty of words

$$score(\square) = \lambda_1 \text{query terms} + \lambda_2 \sum p(\text{names}|e) + \lambda_3 p(\text{entity link to } e|d) \lambda_4 KL\left(p(\text{terms}|e) \parallel p(\text{terms}|d)\right)$$

Neural: Word-Entity Duet Model [Xiong et al, 17]

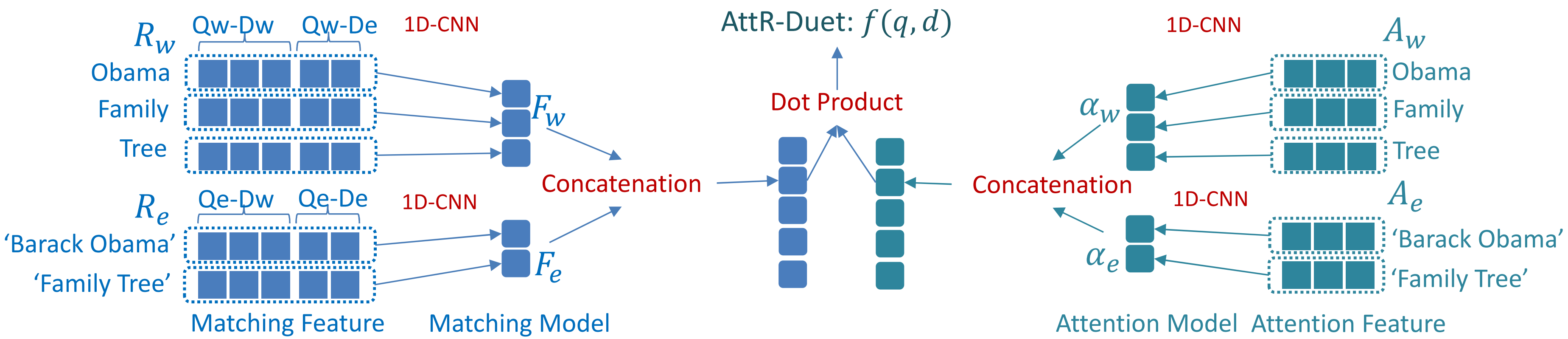


Figure 1: The Architecture of the Attention based Ranking Model for Word-Entity Duet (`AttR-Duet`). The left side models the query-document matching in the word-entity duet. The right side models the importances of query entities using attention features. They together produce the final ranking score.

Image credit: Xiong

Entity Aspects

1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

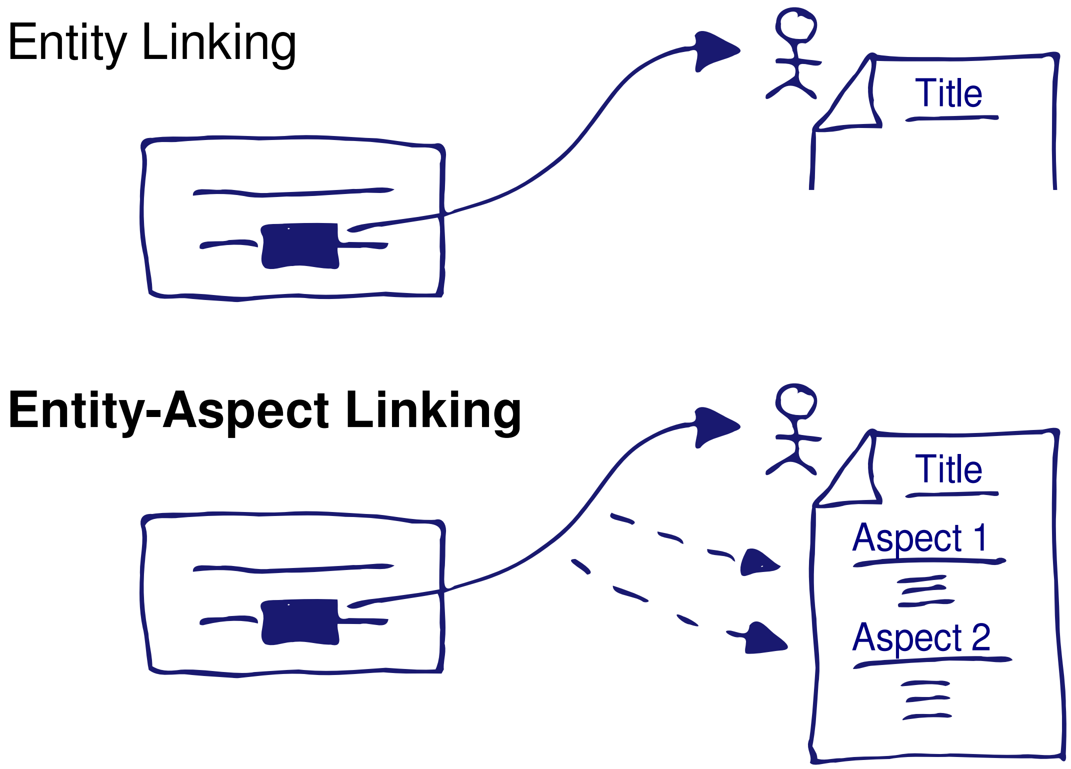
Entity Aspects

Danger: An entity is relevant, but:
only because of one aspect
=> many non-relevant aspects of relevant entities.

Example aspects about UK:
- still a member of the European Union
- is a constitutional monarchy
- the Raspberry Pi was invented in the UK
- there are many great UK bands

Depending on query, some are relevant, some not.

Refining Entity Links with Entity Aspects [Nanni et al, 18]



Here: Using sections from Entity's Wikipedia article as a canonical set of entity aspects.

On data from TREC CAR: P@1 = 68%

Entity Aspects: Sub-topic Classification [Nanni et al, 18]

Results on BreXerch [Zhang et al 17]
(Tweet classification)

Topic	# Tweets
Economy	155
Immigration	52
Sovereignty and influence	50
Security, law enforcement and defense	3
Risk to the Unity of the United Kingdom	30
Transatlantic Trade and Investment Partnership	5
Enlargement of the European Union	12
Proposed consequences of a vote to leave	65
Total	372
Excluded	
General	270
Out-of-topic	108

topic-independent training with L2R!

	P@1
random baseline	0.12
Ranking Approaches	
Content - BM25	0.37
Content - w-emb (cs)	0.36
AL (ours)	**0.43**
Classification Approaches	
Naive Bayes (tf-idf)	0.27
SVM (tf-idf)	0.27
Naive Bayes (w-emb)	0.38
SVM (w-emb)	0.37

Brexit topics

Results

taken from en.wikipedia.org/wiki/Issues_in_the_United_Kingdom_European_Union_membership_referendum,_2016

Conclusion

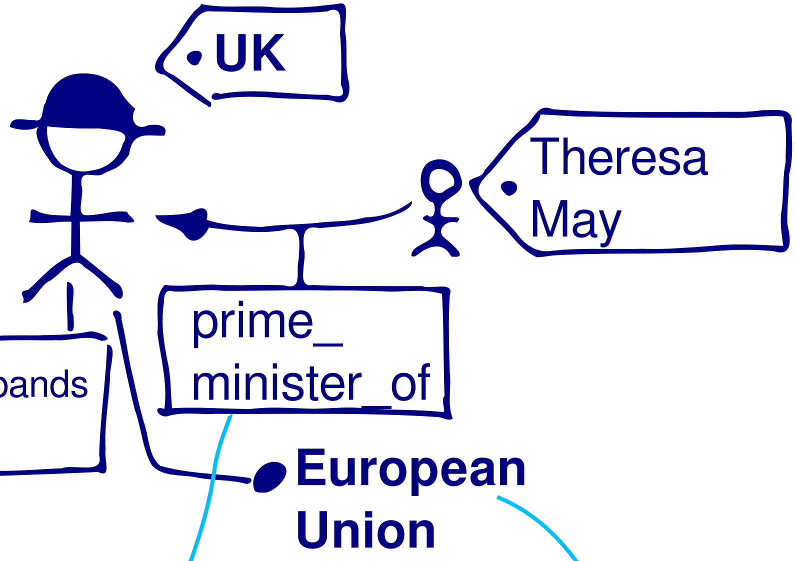
1. Matching entities in documents
2. Find relevant entities
3. Entity types
4. Graph expansion
5. Combination of multiple sources
6. Machine learning
7. Entity aspects

How to Represent Different Knowledge "Units"

As terms?	UK bands
	brexit
As types?	UK member of "European Union"
As is-a?	UK as a European country
Related entities?	`UK` `Theresa_May`
Relations?	`Theresa_May`
	prime_minister_of `UK`
Language Model	p(brexit)=0.4
	p(leave)=0.25
	p(immigration)=0.10

[Reinanda SIGIR15, Liu IRJ15, Prasojo CIKM15]

KG-aware Text Retrieval: Knowledge "Units"

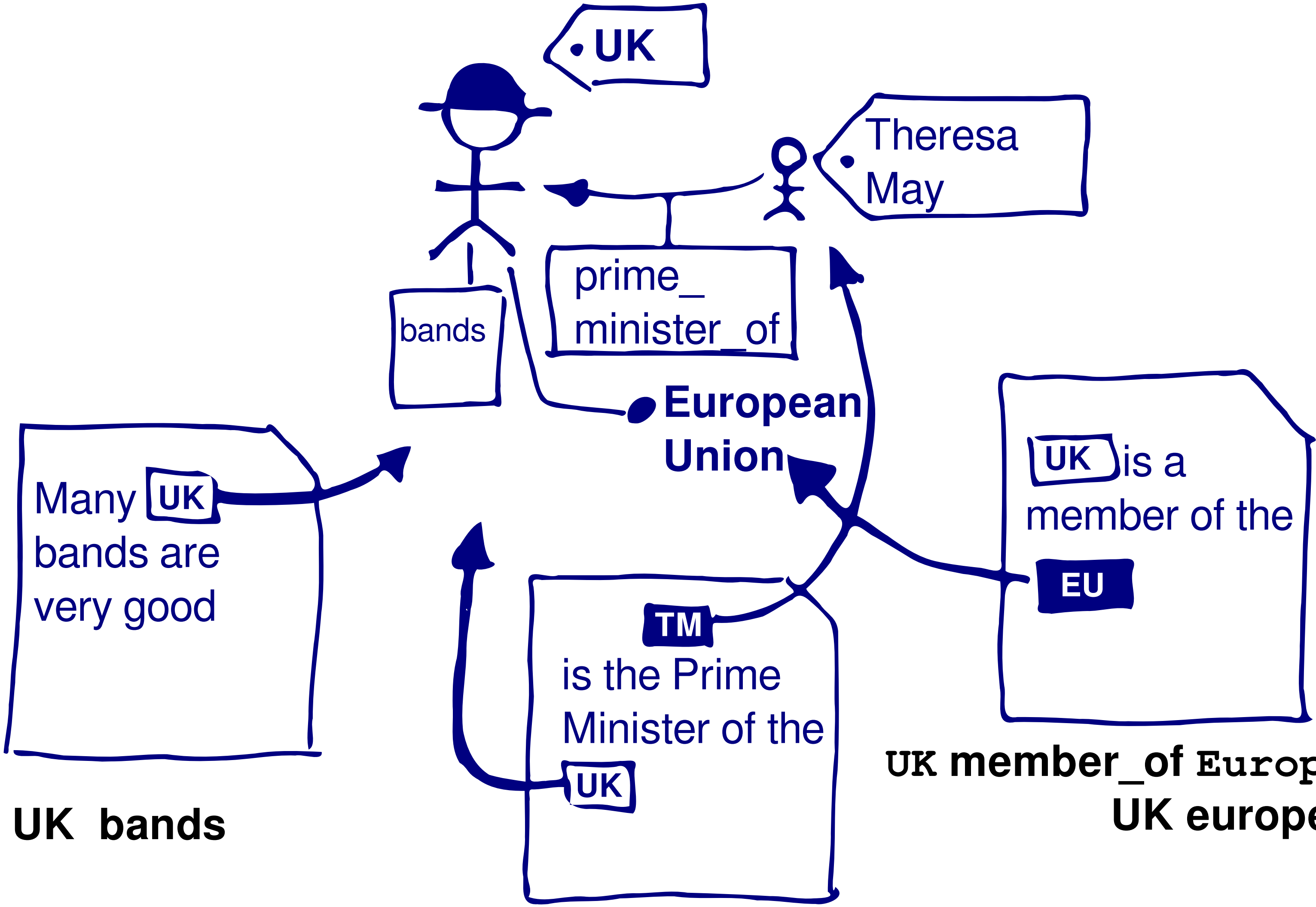


UK bands

UK **member_of** European Union
UK europe

UK Theresa_May
Theresa_May **prime_minister_of** UK

KG-aware Text Retrieval: Knowledge "Units"



UK bands

UK member_of European_Union
UK europe

UK Theresa_May
Theresa_May prime_minister_of UK

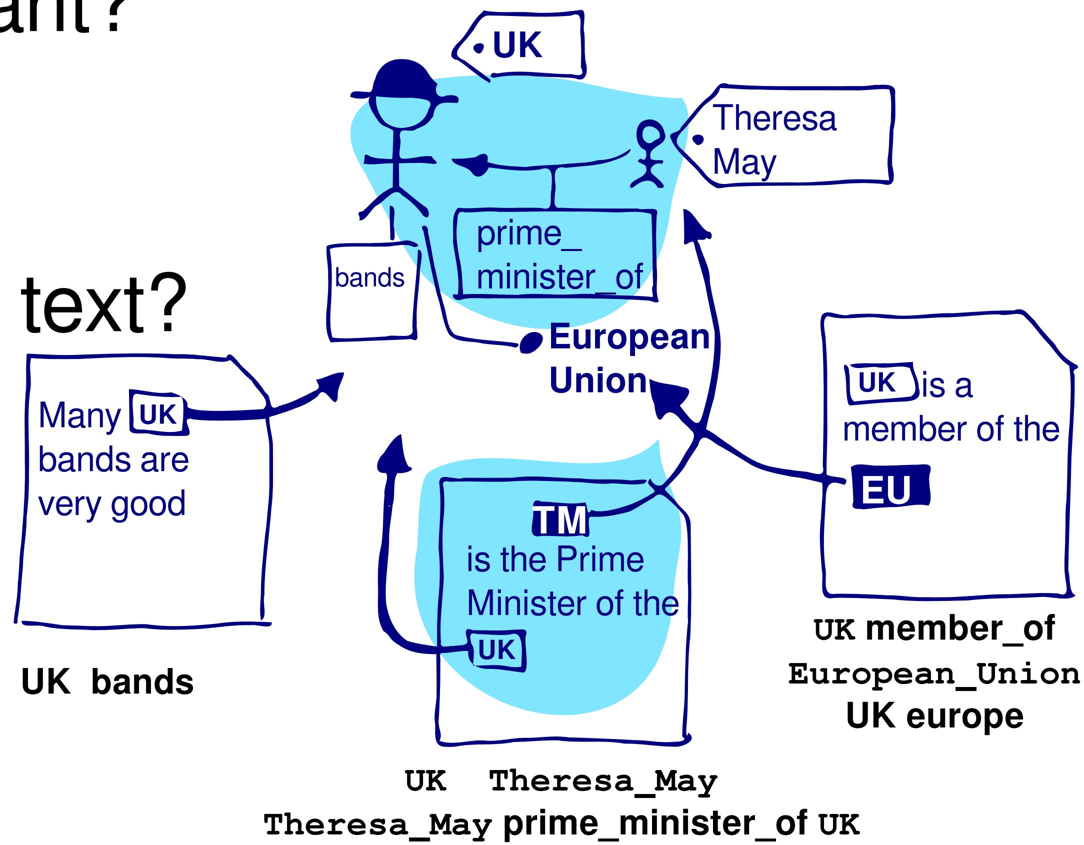
Knowledge "Units": Infer Relevance, Match, Extract

1) Relevance:
Which units are relevant?

2) Match:
How to match units in text?

pseudo relevance feedback

inverse tasks



3) Extract:
How to extract new units? (KB population)

Summary (Part 4)

1. Matching entities in documents
2. Find relevant entities
3. Graph expansion
4. Entity types
5. Combination of multiple sources
6. Machine learning
7. Entity aspects