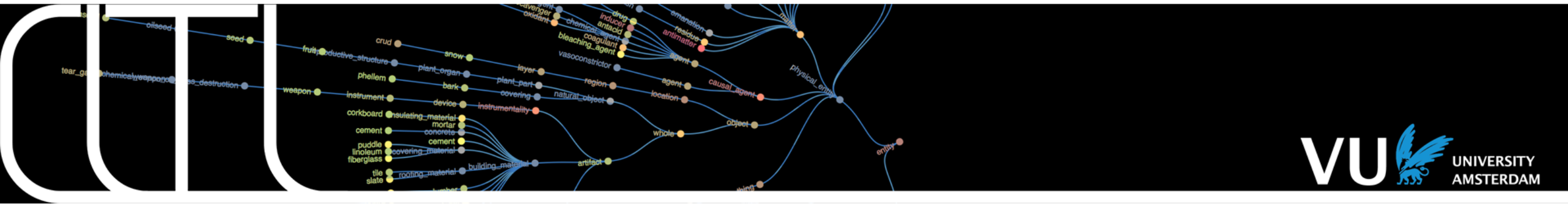
Text Mining CBS

Lecture 5: Named entity disambiguation and linking

Piek Vossen



VU UNIVERSITY AMSTERDAM

Lecture overview

1. Task introduction
    a. Definition
    b. Knowledge bases
    c. Opportunities and challenges
2. Phases of entity linking
3. Entity linkers
    a. Approaches
    b. Tools
4. Evaluation
    a. Aggregation
    b. Example

Entity tasks in NLP

- NER (Recognition): detecting the phrase that is the name of an entity

- NEC (Classification): assigning an entity type to the phrase

- NEL (Linking): establishing the identity of the entity in a given reference database (Wikipedia, DBpedia, YAGO)

- Coreference: any phrase that makes reference to an entity instance, including pronouns, noun phrases, abbreviations, acronyms, etc…

Task definition

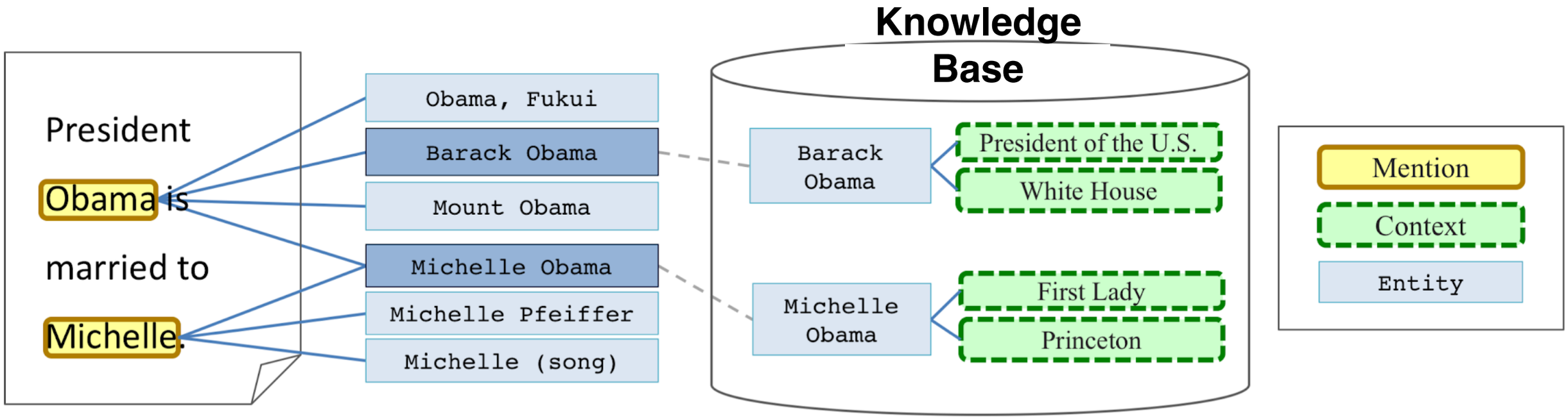
- Potentially ambiguous **entity mention** (“Paris”) needs to be linked to a canonical identifier/**instance** (http://dbpedia.org/resource/Paris) that fits the intended referent in the context of the text

- We find these instances in a **knowledge base**.

Example

“President **Obama** is married to **Michelle**.”

Example


President
Obama is
married to
Michelle.
Obama, Fukui
Barack Obama
Mount Obama
Michelle Obama
Michelle Pfeiffer
Michelle (song)
Knowledge Base
Barack Obama
President of the U.S.
White House
Michelle Obama
First Lady
Princeton
Mention
Context
Entity

Knowledge base

A catalog of things, usually entities. Each one has:

- **a unique identifier or record number, possibly a Unique Resource Identifier (URI):**
  - https://www.wikidata.org/wiki/Q513, https://en.wikipedia.org/wiki/Mount_Everest
- **one or more names**
  - **Mount Everest** -> “Mount Everest”, “Everest”, “Mount Qomolangma”, “Mt. Qomolangma”, “Mount Sagarmatha”, “Qomolangma”, “Chomolangma”, “Mt. Everest”, ...
- **other attributes**
  - Elevation: 8,848m
  - Coordinates: 27°59'17"N, 86°55'31"E
- **Connections to other entities**
  - Continent: Asia
  - Country: China, Country: Nepal
- **Textual description**
  - **example**

Usually a knowledge base has some of these aspects, but not all.

Knowledge base types

The knowledge bases can be classified into two types:

- Unstructured (e.g., Wikipedia)
    - Mostly contains a textual (“unstructured”) description
- Structured (e.g., Wikidata, DBpedia, ...)
    - Contains structured description of an entity
    - Property-value pairs

Unstructured knowledge bases: Wikipedia

https://en.wikipedia.org/wiki/Mount_Everest

Mount Everest

From Wikipedia, the free encyclopedia

Coordinates: 27°59′17″N 86°55′31″E

"Everest" redirects here. For other uses, see Everest (disambiguation).

This article's **tone or style may not reflect the encyclopedic tone used on Wikipedia**. See Wikipedia's guide to writing better articles for suggestions. *(October 2017) (Learn how and when to remove this template message)*

Mount Everest, known in Nepali as **Sagarmatha** (सगरमाथा) and in Tibetan as **Chomolungma** (ཇོ་མོ་གླང་མ), is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. The international border between Nepal (Province No. 1) and China (Tibet Autonomous Region) runs across its summit point.

The current official elevation of 8,848 m (29,029 ft), recognized by China and Nepal, was established by a 1955 Indian survey and subsequently confirmed by a Chinese survey in 1975.[1] In 2005, China remeasured the rock height of the mountain, with a result of 8844.43 m (29,017 ft). There followed an argument between China and Nepal as to whether the official height should be the rock height (8,844 m., China) or the snow height (8,848 m., Nepal). In 2010, an agreement was reached by both sides that the height of Everest is 8,848 m, and Nepal recognizes China's claim that the rock height of Everest is 8,844 m.[5]

In 1865, Everest was given its official English name by the Royal Geographical Society, upon a recommendation by Andrew Waugh, the British Surveyor General of India. As there appeared to be several different local names, Waugh chose to name the mountain after his predecessor in the post, Sir George Everest, despite Everest's objections.[6]

Mount Everest attracts many climbers, some of them highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the "standard route") and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as significant hazards from avalanches and the Khumbu Icefall. As of 2017, nearly 300 people have died on Everest, many of whose bodies remain on the mountain.[7]

The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners into the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). Seven porters were killed in an avalanche on the descent from the North Col. The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether or not they were the first to reach the top. They had been spotted high on the mountain that day but disappeared in the clouds, never to be seen again, until Mallory's body was found in 1999 at 8,155 m (26,755 ft) on the north face. Tenzing Norgay and Edmund Hillary made the first official ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gonpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960.[8][9]

Contents [hide]

Mount Everest

Mount Everest as viewed from Kalapatthar.

Highest point	
Elevation	8,848 metres (29,029 ft)[1] Ranked 1st
Prominence	Ranked 1st (Notice special definition for Everest)
Listing	Seven Summits Eight-thousander Country high point Ultra
Coordinates	27°59′17″N 86°55′31″E[2]
Naming	

Structured knowledge bases: DBpedia & Wikidata

dbpedia.org/page/Mount_Everest

DBpedia | Browse using | Formats

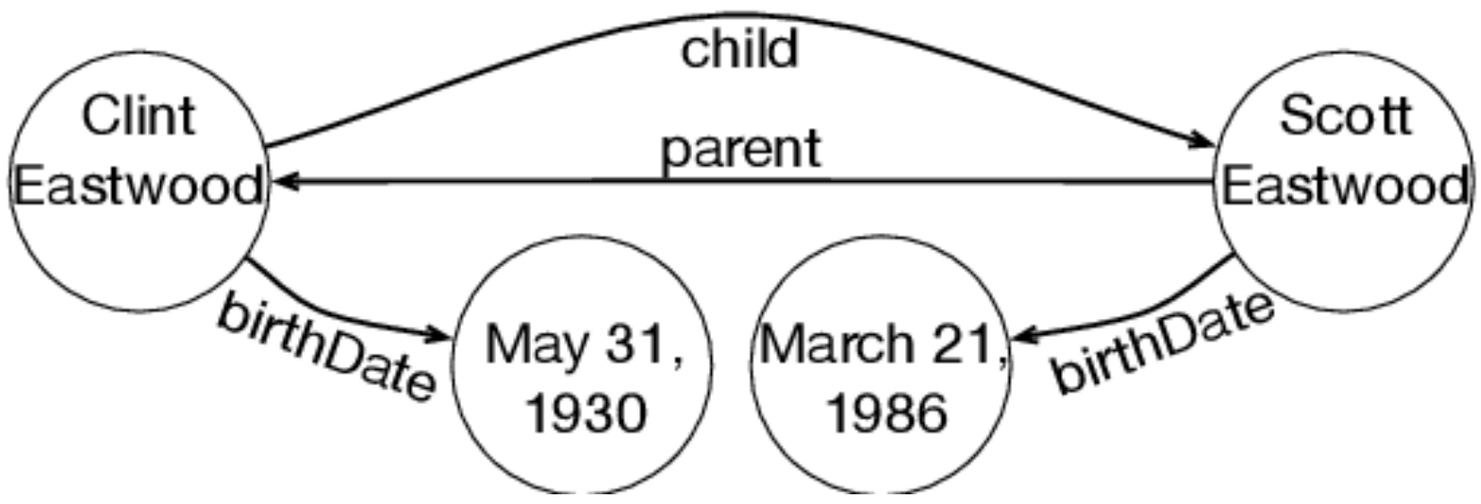
geo:geometry	POINT(86.925277709961 27.988056182861)
geo:lat	27.988056 (xsd:float)
geo:long	86.925278 (xsd:float)
prov:wasDerivedFrom	wikipedia-en:Mount_Everest?oldid=744845387
foaf:depiction	wiki-commons:Special:FilePath/Mount-Everest.jpg
foaf:isPrimaryTopicOf	wikipedia-en:Mount_Everest
foaf:name	Mount Everest (en)
is dbo:deathPlace of	dbr:Shailendra_Kumar_Upadhyaya dbr:Mick_Burke_(mountaineer) dbr:Ray_Genet dbr:Scott_Fischer dbr:Karl_Gordon_Henize dbr:Hristo_Prodanov dbr:David_Sharp_(mountaineer) dbr:Rob_Hall dbr:Pasang_Lhamu_Sherpa dbr:Zygmunt_Andrzej_Heinrich dbr:Mohammad_Khaled_Hossain dbr:Andrew_Irvine_(mountaineer) dbr:Maurice_Wilson

https://www.wikidata.org/wiki/Q513

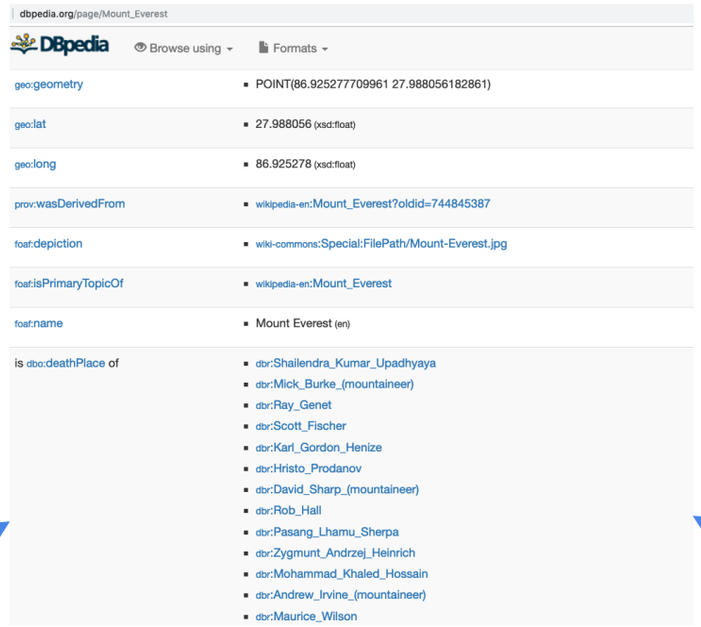
instance of	mountain 0 references
part of	Seven Summits 1 reference Himalayas 1 reference
image	 Mount Everest from Rongbuk may 2005.JPG 3,008 × 2,000; 1.12 MB 0 references

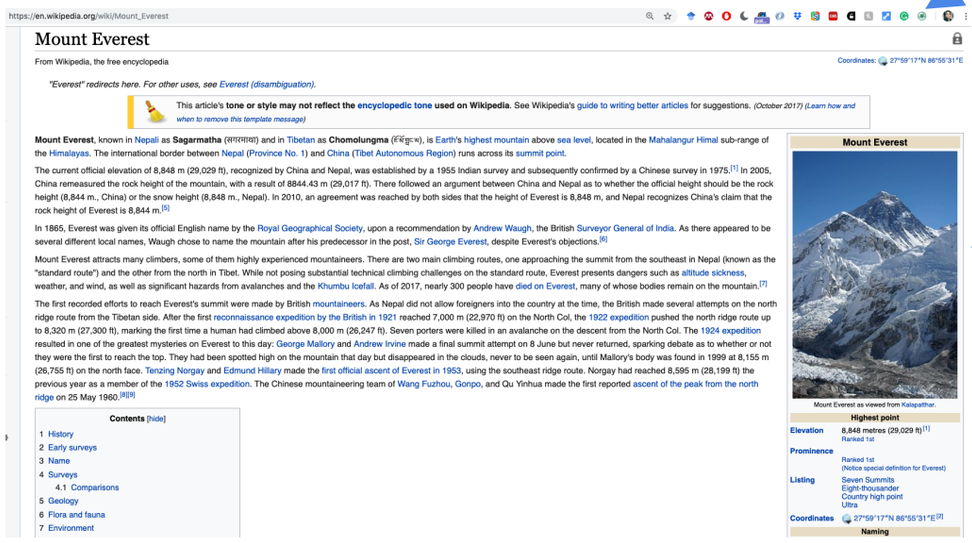
Structured KBs are essentially graph networks

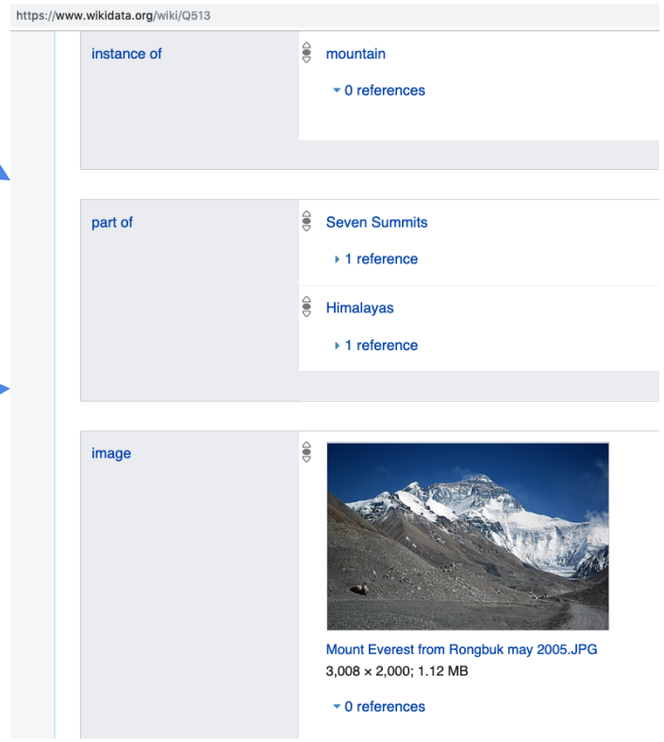
... with billions (!) of such facts



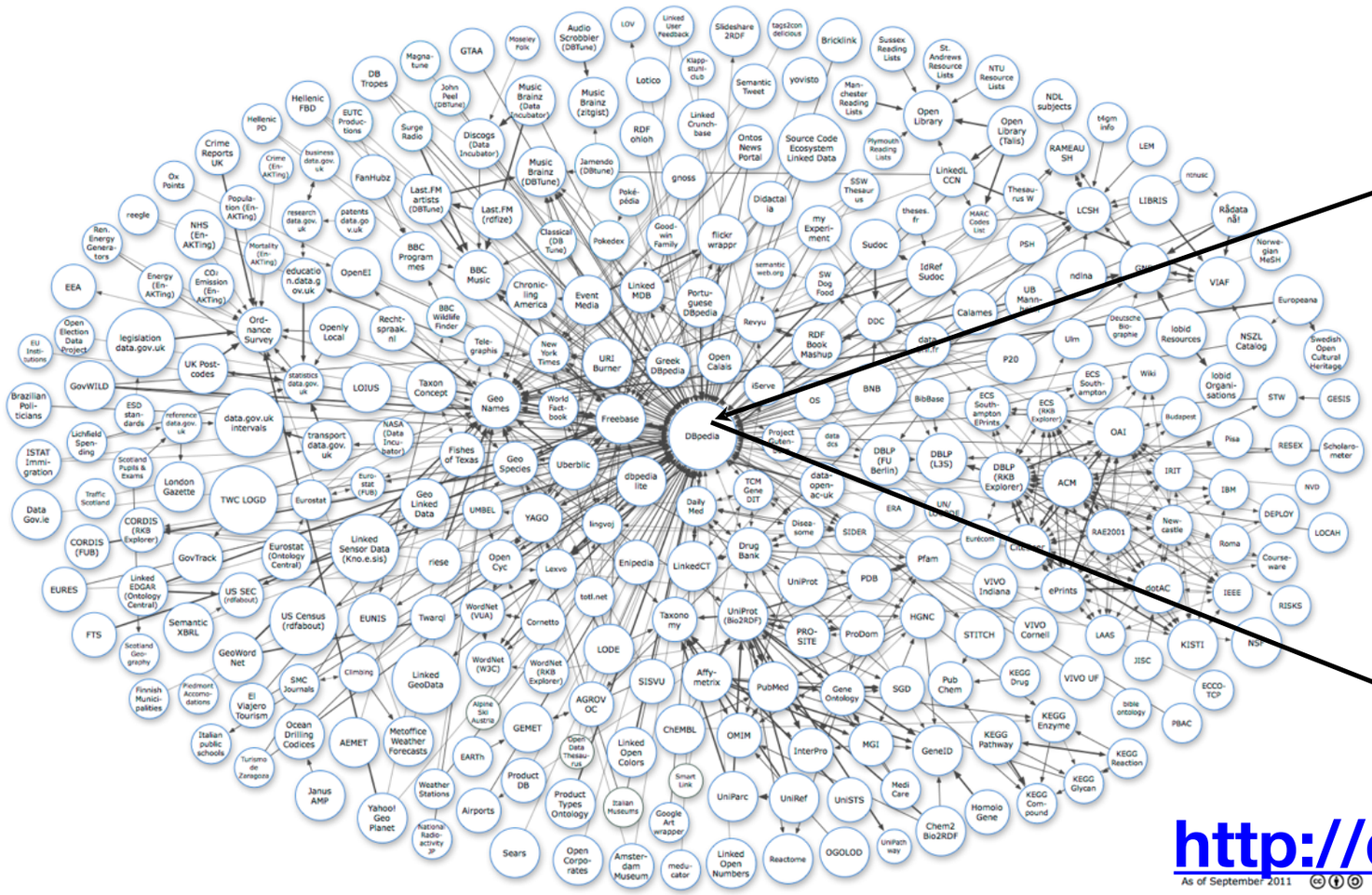
Knowledge bases are also connected to each other







Many more KBs exist, and they are connected -> the LOD Cloud



“John Travolta”

http://dbpedia.org/resource/John_Travolta

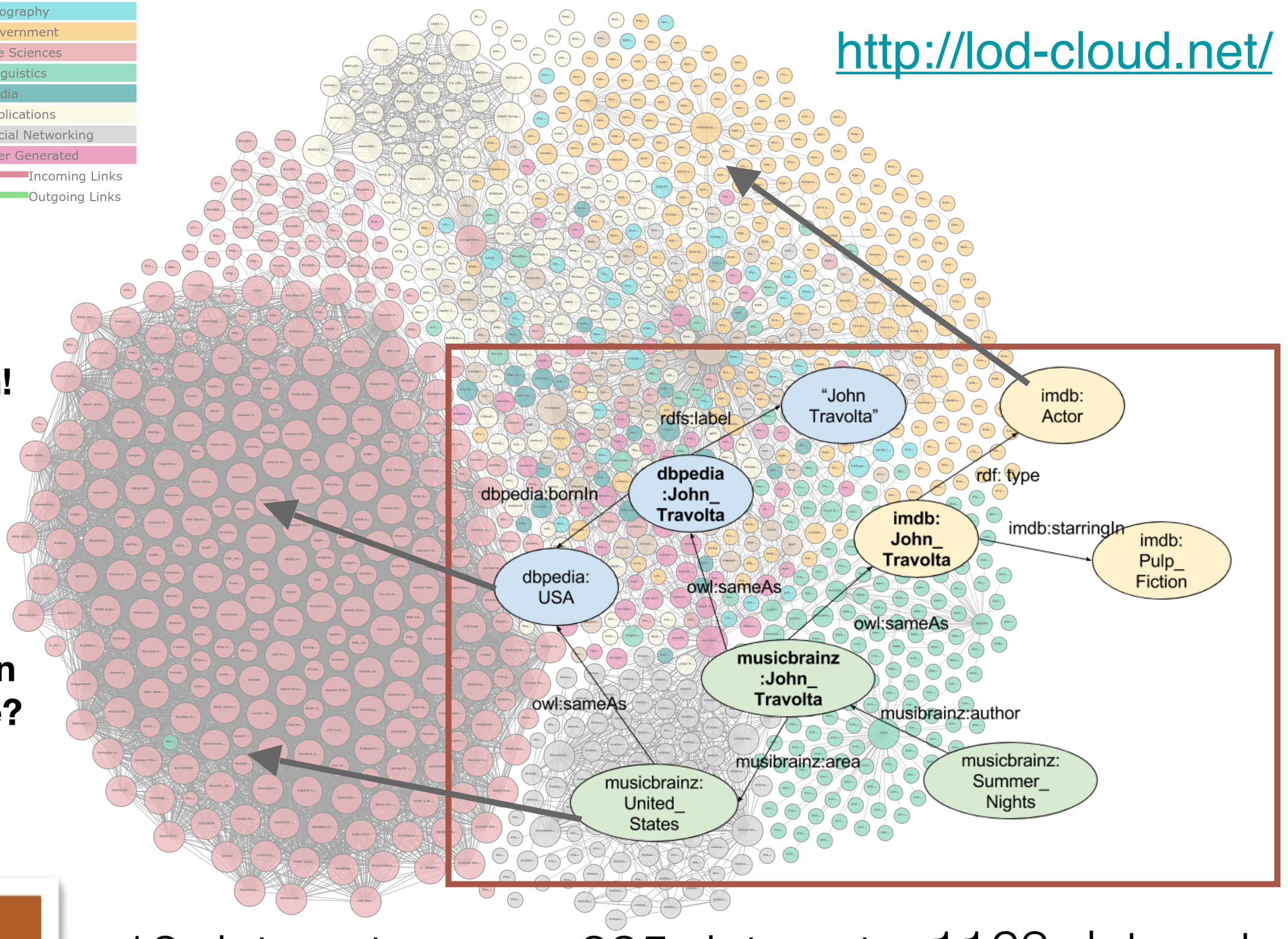
Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated
- Incoming Links
- Outgoing Links

http://lod-cloud.net/

I love John!

Which John do you love?



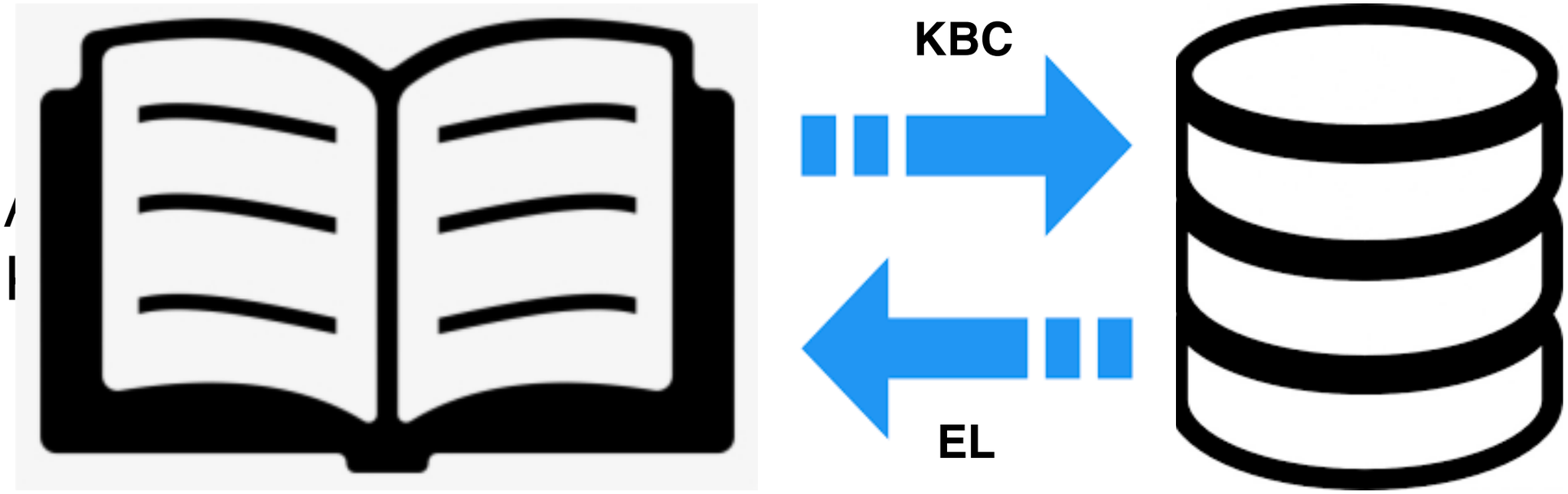
Power of the division of labour

12 data sets
2 billion triples
2007

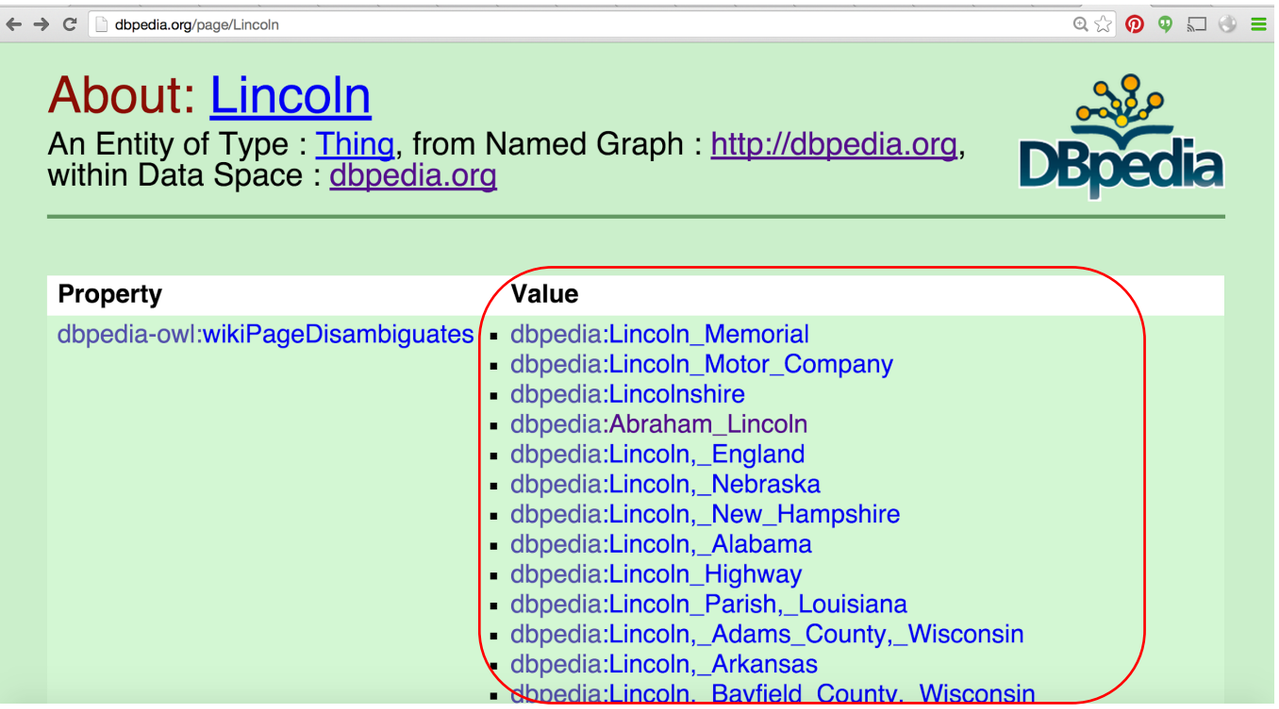
295 data sets
31 billion triples
2011

1163 data sets
2017

Other benefits of connecting text and knowledge bases



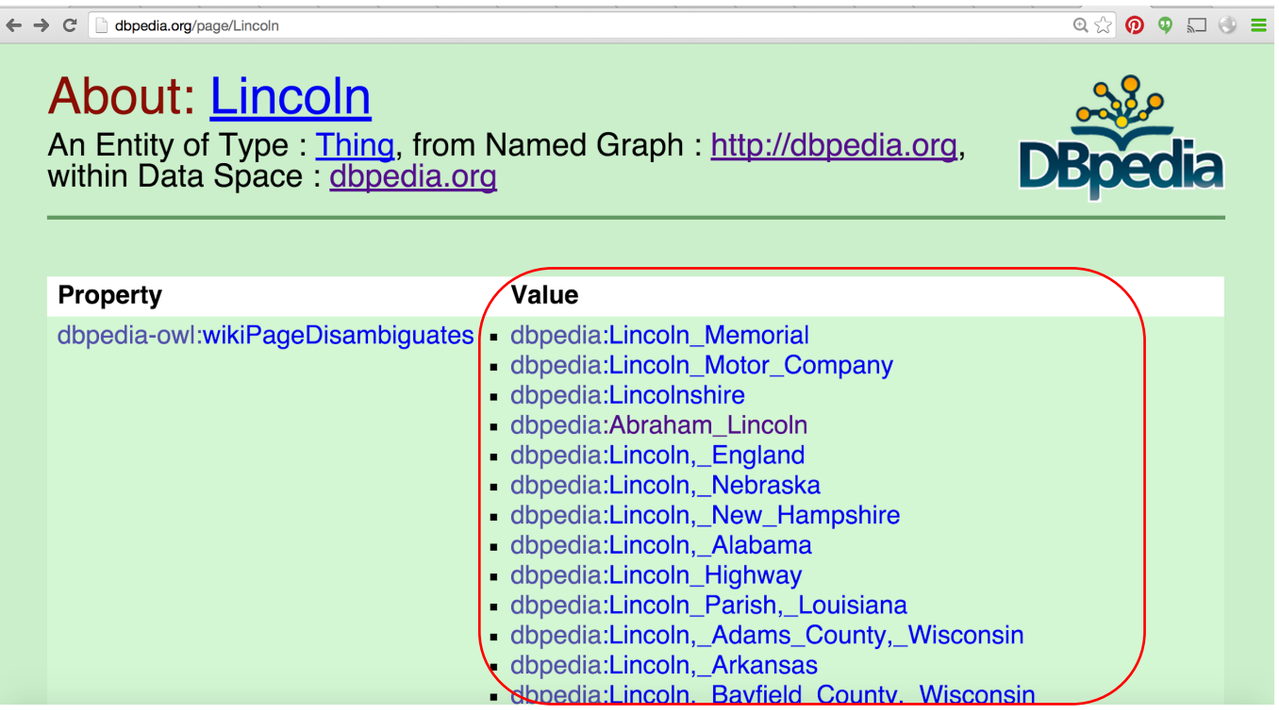
Named Entity Disambiguation isn't that easy though

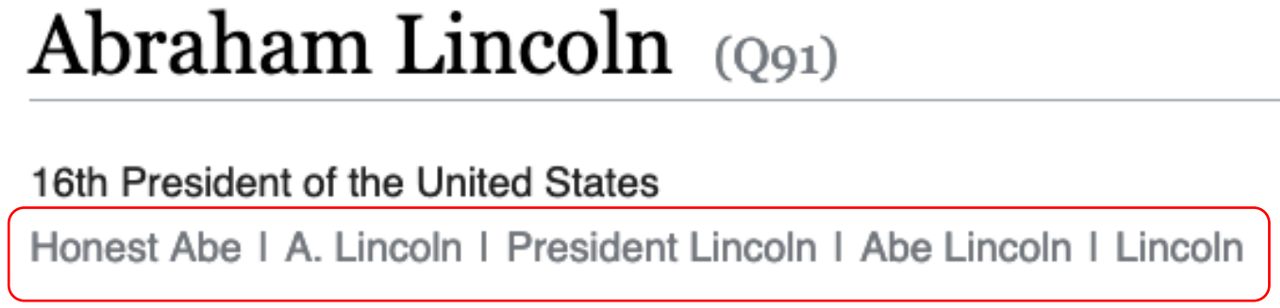


a. name **ambiguity**

Very frequent => Wikipedia and DBpedia have special **disambiguation** pages that list the entities that are referred to by a mention.

Named Entity Disambiguation isn't that easy though





a. name **ambiguity**

Very frequent => Wikipedia and DBpedia have special **disambiguation** pages that list the entities that are referred to by a mention.

b. name variation

(+ "Mr. Lincoln", "Lincoln", "Abraham", "Abe", "Honest Abe")

The meaning puzzle: Who is Ford?

- President Woodrow Wilson asked ***Ford***[?] to run as a Democrat for the United States Senate from Michigan in 1918.

- Gerald Ford
- Ford, the motor company
- Henry Ford

The meaning puzzle: Who is Ford?

President [6] Woodrow Wilson [10] asked [7] Ford [8] to run [41] as a Democrat [2] for the United States [4] Senate [2] from Michigan [3] in 1918.

6*10*7*8*41*2*4*2*3= 6,612,480 combinations of word senses and entities

The meaning puzzle: Who is Ford?

- President Woodrow Wilson asked ***Ford***[?] ***to run as*** a Democrat for the United States Senate from Michigan in 1918.
- Semantic knowledge:
  - meaning of ***to run as***
- World knowledge:
  - run as senator: +human, >18 years old, US citizen
  - Gerald Ford born in 1917, so 1 year old
  - Henry Ford the founder of Ford Motor Company, born in 1863, died in 1947, so 56 years old

c. Missing (NIL) entities

SAN BENITO – Police have released more information connected to a murder investigation in San Benito.

Edgar Gonzalez, 30, was shot and killed last Thursday.

It happened on Buena Vida Street near the expressway and Sam Houston.

Police released a photo of a white Chevrolet Tahoe they say was used as a getaway vehicle.

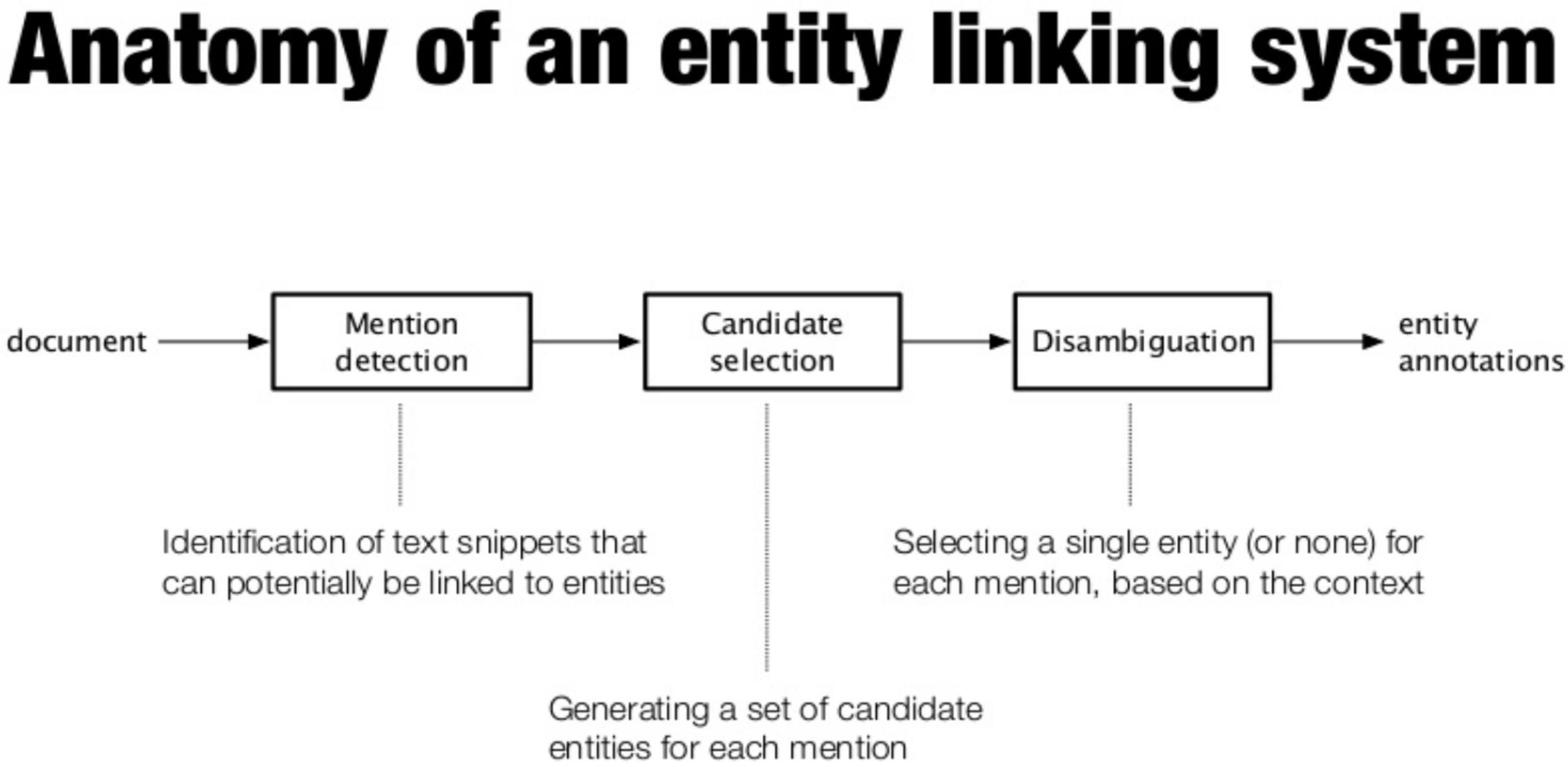
Police are asking nearby businesses to check their surveillance video for any images of this vehicle.

If you have any information, contact the San Benito Police Department at 361-3880.

What to link “Edgar Gonzalez” to???

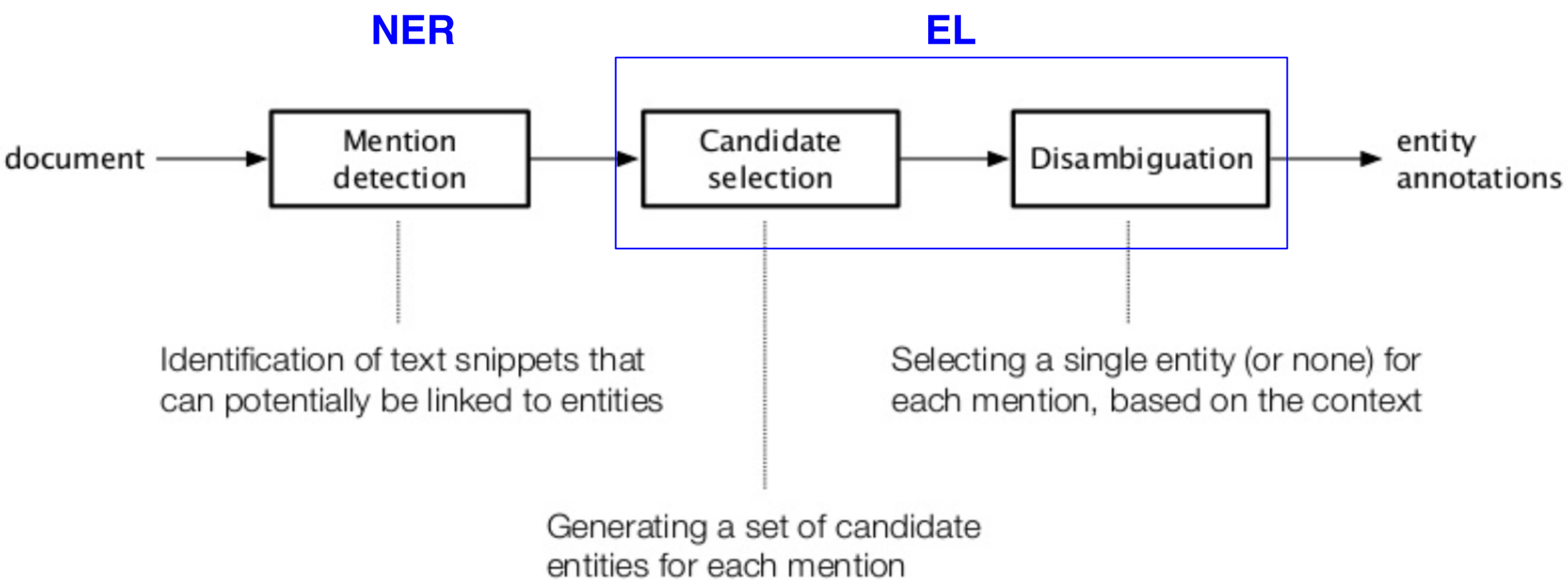
https://www.krgv.com/news/police-seeking-surveillance-footage-to-aid-in-san-benito-murder-investigation

2. Components of an entity linking system



Source: https://www.slideshare.net/krisztianbalog/entity-linking-65308055

Anatomy of an entity linking system



Source: https://www.slideshare.net/krisztianbalog/entity-linking-65308055

Phase I: Recognition

Same as we saw in the previous lecture

Detect entity mentions in text

Phase II: Candidate generation/selection

- For each of the recognised mentions in text, get the potential referents (instances) in a knowledge base (KB), following the “closed world assumption” (=the world is in the KB).
- The goal is to balance between generating too many candidates (too much ‘noise’) and generating too little candidates (missing the correct one)
- Trade-off between precision and recall
- Candidate generation is an art by itself!

But... how do you choose the top X (or 30) candidates?

- We need a way to rank them somehow.
- A common ranking criteria is **commonness**: for a given mention, how relatively often it refers to some instance in Wikipedia.
- For example, of all the mentions of “Germany” in Wikipedia, what is the percentage that refers to the country vs the football club vs the handball club vs the government vs etc.

- Perform the ranking of candidate entities based on their overall popularity, i.e., "most common sense"

$$P(e|m) = \frac{n(m,e)}{\sum_{e'} n(m,e')}$$

$n(m,e)$ → the number of times entity e is the link destination of mention m

$\sum_{e'} n(m,e')$ → total number of times mention m appears as a link

Example

Entity	Commonness
FIFA_World_Cup	0.2358
FIS_Apline_Ski_World_Cup	0.0682
2009_FINA_Swimming_World_Cup	0.0633
World_Cup_(men's_golf)	0.0622
...	

Bulgaria's best **World Cup** performance was in the **1994 World Cup** where they beat **Germany**, to reach the semi-finals, losing to Italy, and finishing in fourth ...

Entity	Commonness
1998_FIFA_World_Cup	0.9556
1998_IAAF_World_Cup	0.0296
1998_Alpine_Skiing_World_Cup	0.0059
...	

Entity	Commonness
Germany	0.9417
Germany_national_football_team	0.0139
Nazi_Germany	0.0081
German_Empire	0.0065
...	

Also, observe:

- Dominance within a form
- Topical bias

In practice, about 30 candidates per mention is enough.

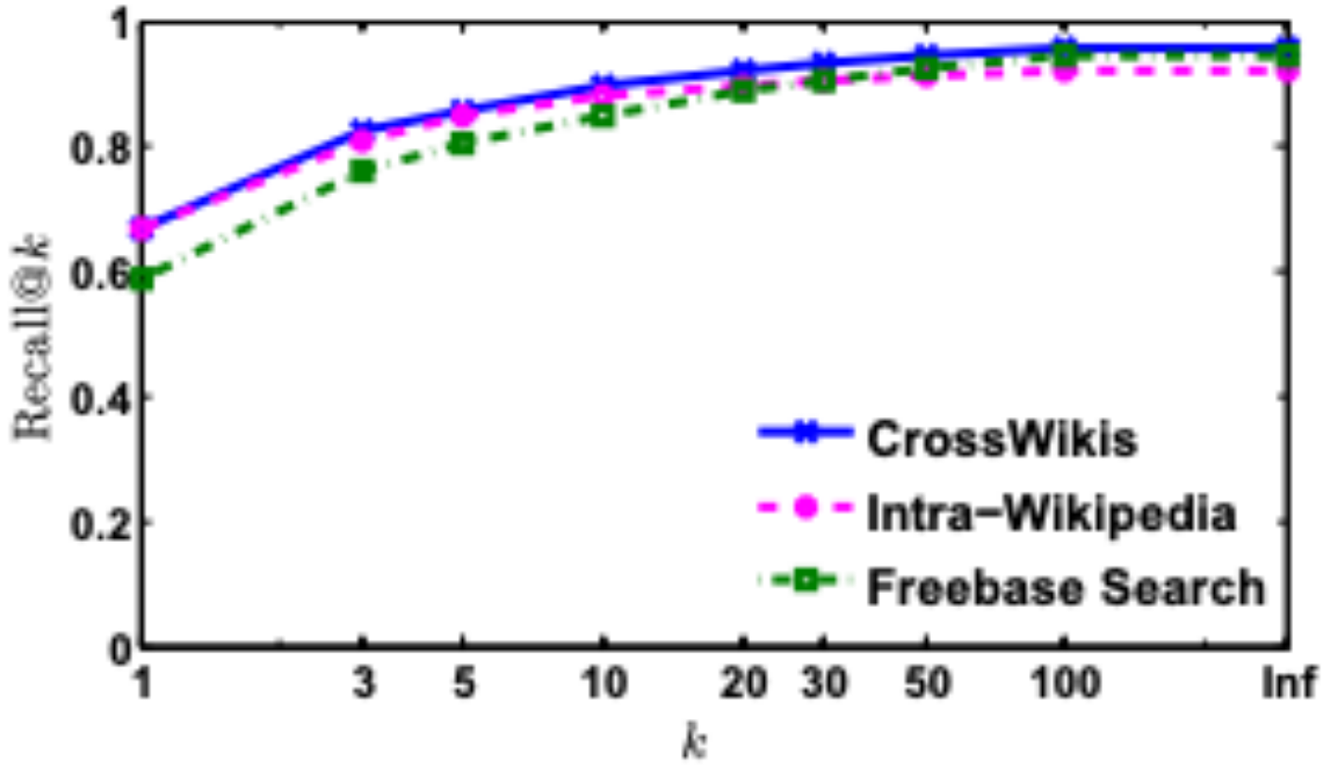


Figure 3: Recall@k on an aggregate of nine data sets, comparing three **candidate generation** methods.

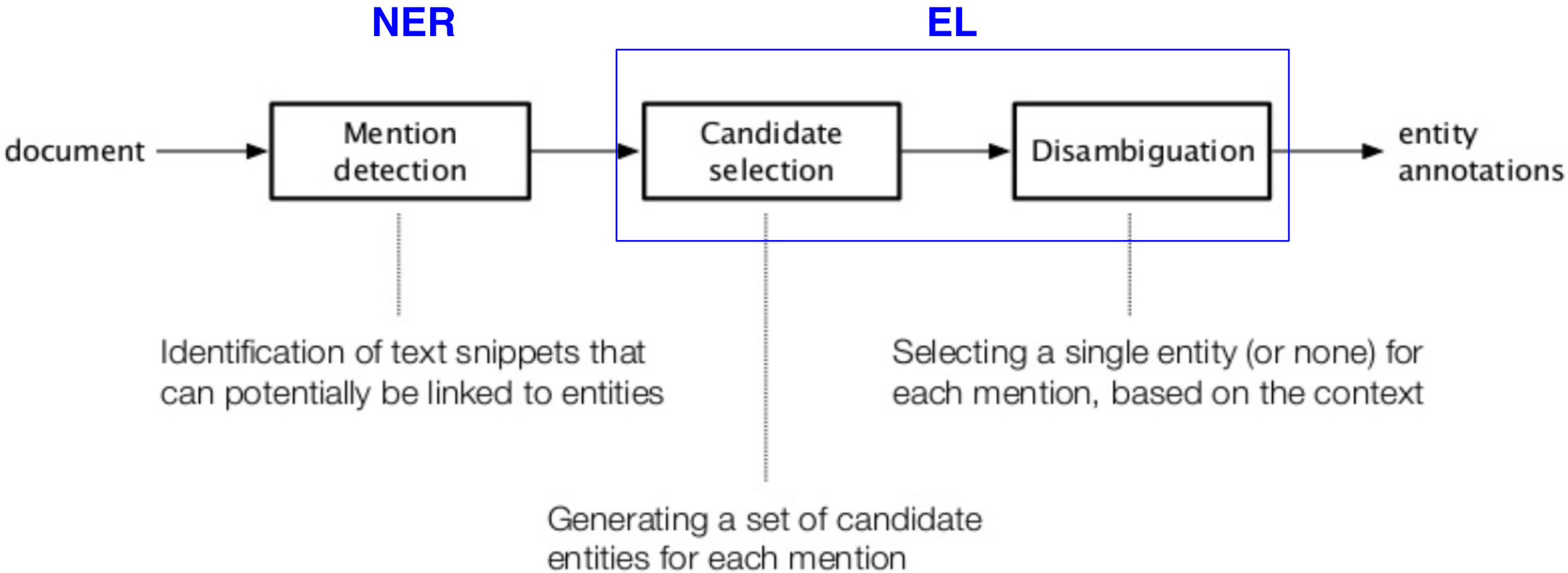
Source: https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00141

Phase III: Disambiguation

Goal: decide which of the candidates (or none) is the correct referent.

When would this phase be easy and when difficult?

3. Tools/systems



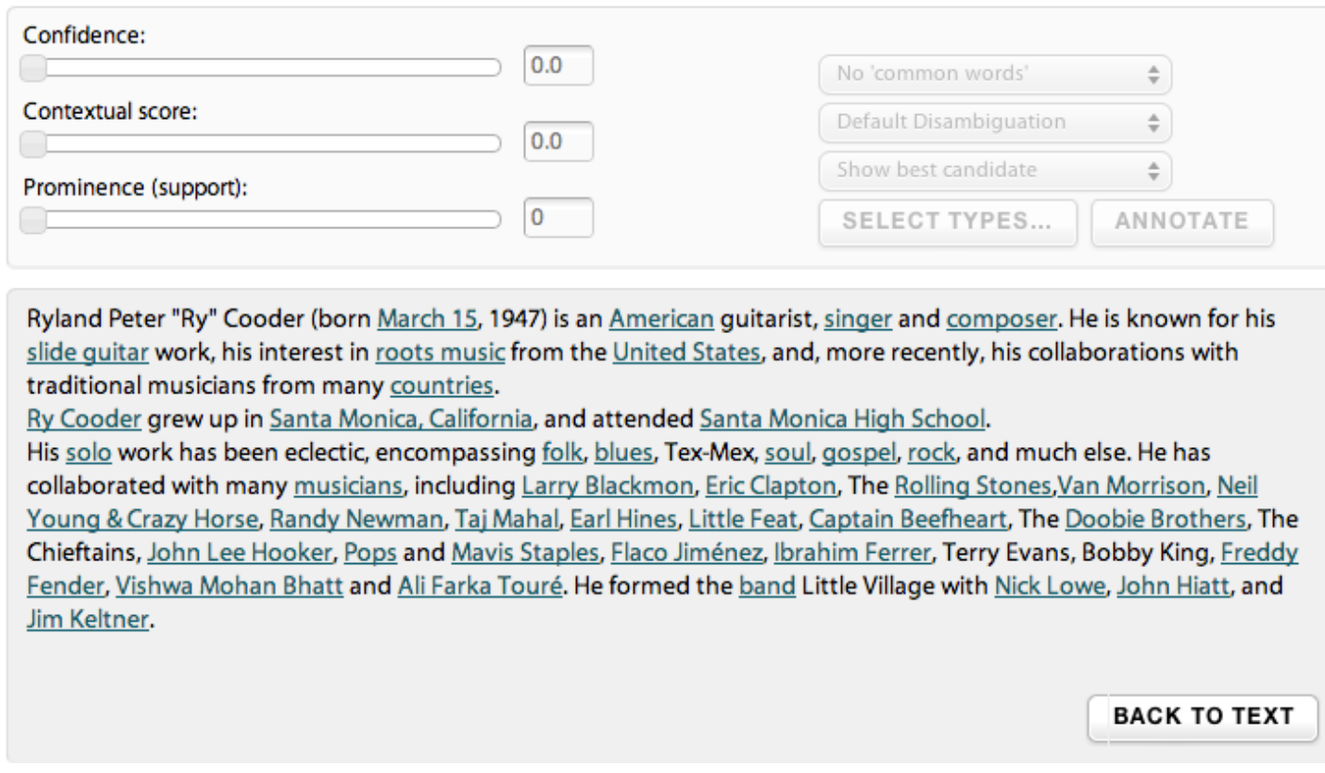
Entity linking methods

	Word-based	**Graph-based**
Main idea	Find the candidate with the most similar description to the one of a mention in text	Find candidates that are coherent with each other according to connections in the KB
Scoring example	measure text similarity, combine with **TF/IDF** weighting to measure relevance of a word	Put all candidates with their facts in a graph network and prune until only one candidate per mention is left
Decision unit	individual/local	collective/global
KB	unstructured (Wikipedia)	Structured (DBpedia, etc.)
Example	DBpedia Spotlight	AIDA/AGDISTIS

TF/IDF = term frequency * inverse document frequency
Measures the degree to which terms are important for a document, based on the frequency in the document but normalised by checking if it occurs in all documents or just a few

3a. Word-based methods: DBpedia Spotlight

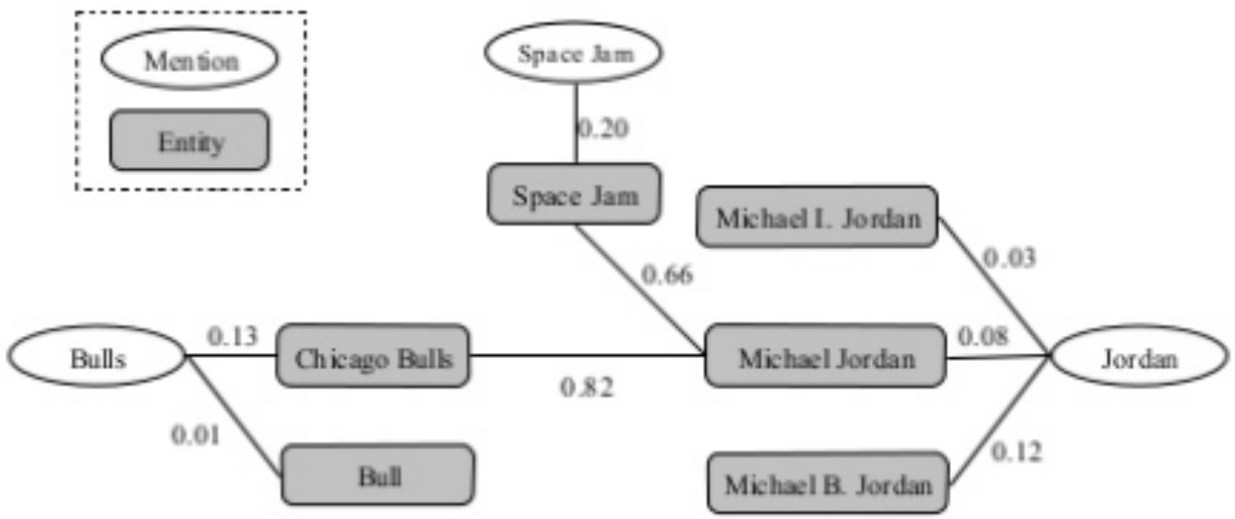
- Compute cosine similarity between the text paragraph with an entity mention and Wikipedia descriptions of each candidate.
- Decide for one mention at a time.
- The linking can be restricted to certain types or even to a custom set of entities.



3b. Graph-based methods: AIDA and AGDISTIS

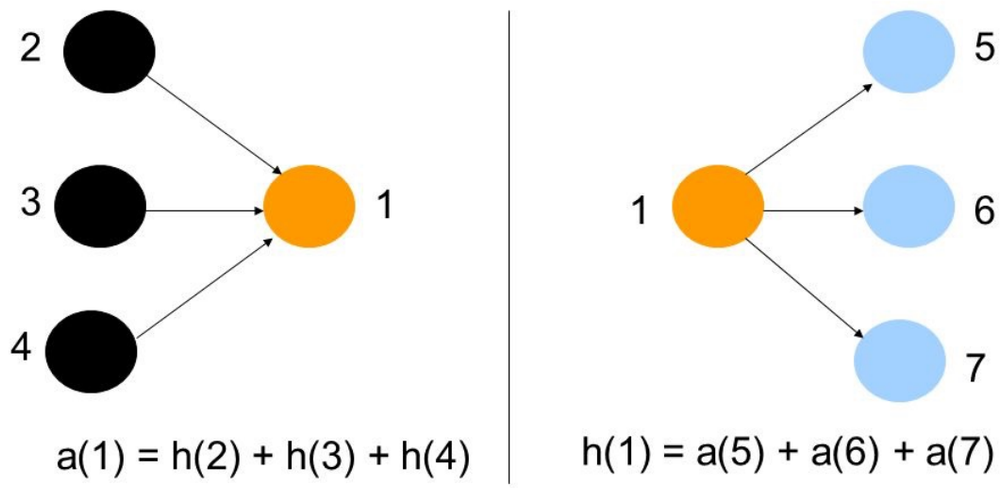
1. Construct a subgraph that contains all entity candidates with some facts from a KB.
2. Find the best connected candidates per mention:

Example



Compute relatedness between the candidates (AIDA)

Authority and Hubness



Find the “hubs” in the graph (AGDISTIS)

Local vs global disambiguation

- Note that the idea in the graph-based approaches is to make the optimal global decision (we disambiguate all entities together).

- This is different than in DBpedia Spotlight, where we disambiguate entities one by one.

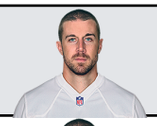
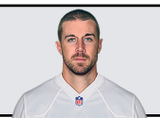
4. Evaluation

- The correctness of an entity linking system is measured in terms of **precision, recall, and F1-score.**
- You already know these metrics, but... here we aggregate the scores differently.
- In Sentiment classification, we compute a score **per class** (positive, neutral, negative).

4. Evaluation

- We could do the same in entity linking, but here we have far too many classes (millions).
- For this reason, we usually evaluate entity linking by aggregating **per mention occurrence.**
- Instead of computing confusion between the classes, here we:
    1. Assign a true positive (TP), false positive (FP), and/or false negative (FN) per mention occurrence
    2. Count the TPs, FPs, and FNs across all mentions
    3. Compute precision, recall, and F1-scores once on top of these

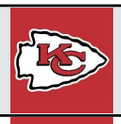
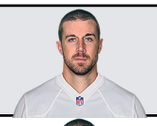
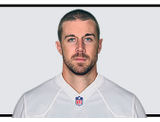
4. EL evaluation example

	Chiefs	Alex Smith	Washington	Smith
GOLD				
SYSTEM				NIL
TP, FP, FN	TP	TP	FP, FN	FN

“It seems like months ago that the Chiefs traded Alex Smith to Washington...”
Smith, 33, originally entered ...”

(https://profootballtalk.nbcsports.com/2018/03/14/washington-announces-alex-smith-trade/)

4. EL evaluation example

	Chiefs	Alex Smith	Washington	Smith
GOLD				
SYSTEM				NIL
TP, FP, FN	TP	TP	FP, FN	FN

“It seems like months ago that the Chiefs traded Alex Smith to Washington...”
Smith, 33, originally entered ...”

```
precision=TP/(TP+FP)=2/3~0.67
recall=TP/(TP+FN)=2/4=0.5
f1=2*precision*recall/(precision+recall)=0.67/1.17=0.57
```

Datasets for Entity Linking

Group	Data Set	# of Mentions	Entity Types	KB	# of NILs	Eval. Metric
UIUC	ACE	244	Any Wikipedia Topic	Wikipedia	0	BOC F1
	MSNBC	654	Any Wikipedia Topic	Wikipedia	0	BOC F1
AIDA	AIDA-dev	5917	PER,ORG,LOC,MISC	Yago	1126	Accuracy
	AIDA-test	5616	PER,ORG,LOC,MISC	Yago	1131	Accuracy
TAC KBP	TAC09	3904	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	2229	Accuracy
	TAC10	2250	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1230	Accuracy
	TAC10T	1500	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	426	Accuracy
	TAC11	2250	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1126	B^3+ F1
	TAC12	2226	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1049	B^3+ F1

- ACE = Automatic content extraction evaluation (follow-up of Message Understanding Competition, MUC), http://www.itl.nist.gov/iad/mig/tests/ace/
- MSNBC = news https://cogcomp.cs.illinois.edu/page/resource_view/4
- AIDA = Accurate Online Disambiguation of Named Entities in Text and Tables, https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/
- TAC = Text Analysis Conference, http://www.nist.gov/tac/
- KBP = Knowledge Base Population, www.nist.gov/tac/2016/KBP/

Vinculum: ablation study

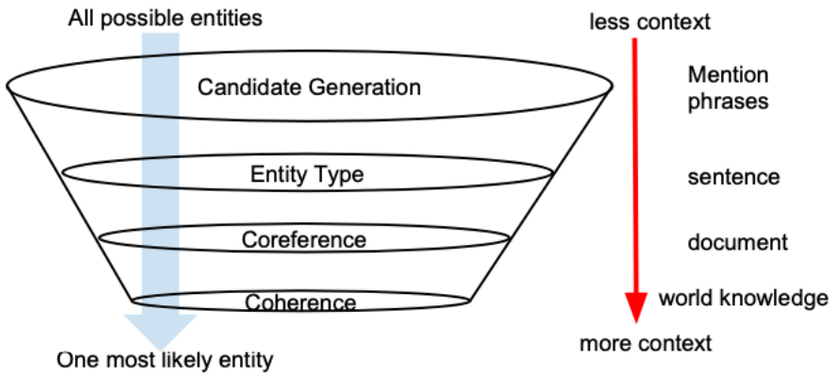


Figure 2: The process of finding the best entity for a mention. All possible entities are sifted through as VINCULUM proceeds at each stage with a widening range of context in consideration.

	ACE		MSNBC		AIDA-dev		AIDA-test	
	R	P	R	P	R	P	R	P
NER	89.7	10.9	77.7	65.5	89.0	75.6	87.1	74.0
+NP	96.0	2.4	90.2	12.4	94.7	21.2	92.2	21.8
+DP	96.8	1.8	90.8	9.3	95.8	14.0	93.8	13.5
+NP+DP	98.0	1.2	92.0	5.8	95.9	9.4	94.1	9.4

Table 3: Performance(%, R: Recall; P: Precision) of the correct mentions using different **mention extraction** strategies. ACE and MSNBC only annotate a subset of all the mentions and therefore the absolute values of precision are largely underestimated.

mention detection

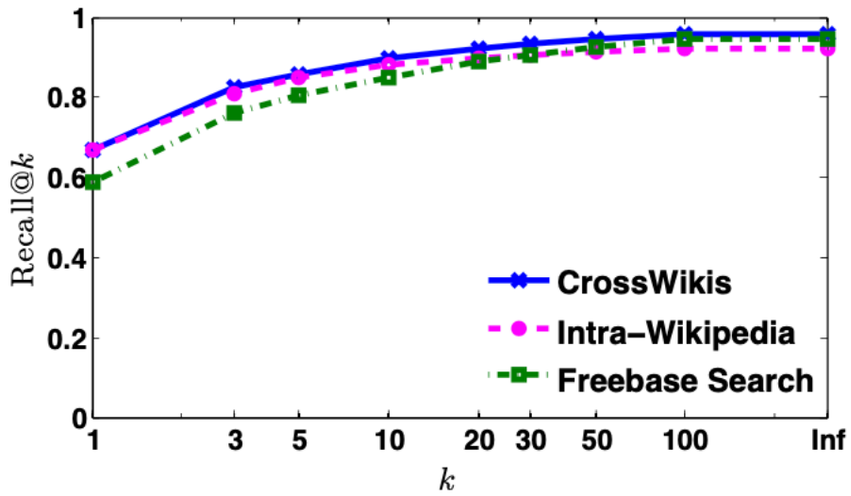


Figure 3: Recall@k on an aggregate of nine data sets, comparing three **candidate generation** methods.

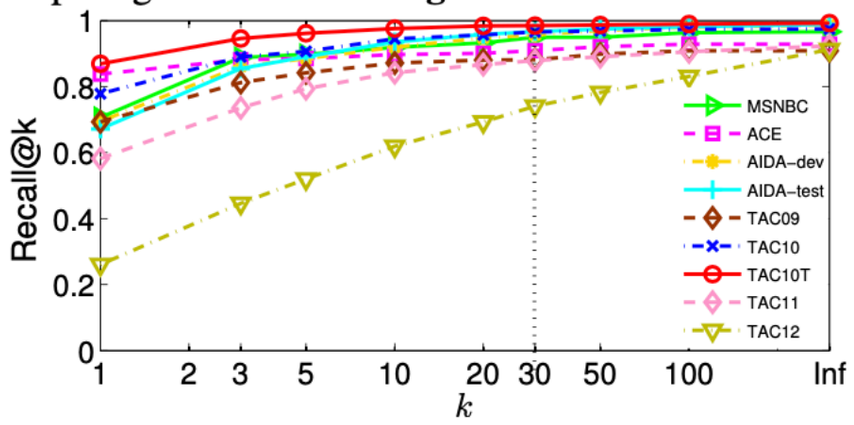


Figure 4: Recall@k using CrossWikis for candidate generation, split by data set. 30 is chosen to be the cut-off value in consideration of both efficiency and accuracy.

candidate generation

System performance

baseline

types

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
CrossWikis only	80.4	85.6	86.9	78.5	62.4	62.6	60.4	87.7	70.3
+NER	79.2	83.3	85.1	76.6	61.1	66.4	**66.2**	77.0	71.8
+FIGER	**81.0**	**86.1**	**86.9**	**78.8**	**63.5**	**66.7**	64.6	**87.7**	**75.4**
+NER(GOLD)	**85.7**	87.4	88.0	80.1	**66.7**	72.6	72.0	89.3	83.3
+FIGER(GOLD)	84.1	**88.8**	**89.0**	**81.6**	66.1	**76.2**	**76.5**	**91.8**	**87.4**

Table 4: Performance (%) after **incorporating entity types**, comparing two sets of entity types (NER and FIGER). Using a set of fine-grained entity types (FIGER) generally achieves better results.

coherence

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
no COH	80.9	86.2	87.0	78.6	59.9	68.9	66.3	87.7	86.6
+NGD	**81.8**	85.7	86.8	79.7	63.2	**69.5**	**67.7**	88.1	86.8
+REL	81.2	86.3	87.0	79.3	63.1	69.1	66.4	**88.5**	86.1
+BOTH	81.4	**86.8**	**87.0**	**79.9**	**63.7**	69.4	67.5	**88.5**	**86.9**

Table 5: Performance (%) after re-ranking candidates using coherence scores, comparing two **coherence measures** (NGD and REL). "no COH": no coherence based re-ranking is used. "+BOTH": an average of two scores is used for re-ranking. Coherence in general helps: a combination of both measures often achieves the best effect and NGD has a slight advantage over REL.

Error analysis

Error Category	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
Metonymy	16.7%	0.0%	3.3%	0.0%	0.0%	60.0%	60.0%	5.3%	20.0%
Wrong Entity Types	13.3%	23.3%	20.0%	6.7%	10.0%	6.7%	10.0%	31.6%	5.0%
Coreference	30.0%	6.7%	20.0%	6.7%	3.3%	0.0%	0.0%	0.0%	20.0%
Context	30.0%	26.7%	26.7%	70.0%	70.0%	13.3%	16.7%	15.8%	15.0%
Specific Labels	6.7%	36.7%	16.7%	10.0%	3.3%	3.3%	3.3%	36.9%	25.0%
Misc	3.3%	6.7%	13.3%	6.7%	13.3%	16.7%	10.0%	10.5%	15.0%
# of examined errors	30	30	30	30	30	30	30	19	20

Table 9: **Error analysis:** We analyze a random sample of 250 of VINCULUM's errors, categorize the errors into six classes, and display the frequencies of each type across the nine datasets.

Category	Example	Gold Label	Prediction
Metonymy	South Africa managed to avoid a fifth successive defeat in 1996 at the hands of the All Blacks ...	South Africa national rugby union team	South Africa
Wrong Entity Types	Instead of Los Angeles International, for example, consider flying into Burbank or John Wayne Airport ...	Bob Hope Airport	Burbank, California
Coreference	It is about his mysterious father, Barack Hussein Obama, an imperious if alluring voice gone distant and then missing.	Barack Obama Sr.	Barack Obama
Context	Scott Walker removed himself from the race, but Green never really stirred the passions of former Walker supporters, nor did he garner outsized support "outstate".	Scott Walker (politician)	Scott Walker (singer)
Specific Labels	What we like would be Seles , (Olympic champion Lindsay) Davenport and Mary Joe Fernandez .	1996 Summer Olympics	Olympic Games
Misc	NEW YORK 1996-12-07	New York City	New York

Table 8: We divide linking errors into **six error categories** and provide an example for each class.

Entity linking pipelines

	VINCULUM	AIDA	WIKIFIER
Mention Extraction	NER	NER	NER, noun phrases
Candidate Generation	**CrossWikis**	an intra-Wikipedia dictionary	an intra-Wikipedia dictionary
Entity Types	**FIGER**	NER	NER
Coreference	find the representative mention	-	re-rank the candidates
Coherence	link-based similarity, relation triples	link-based similarity	link-based similarity, relation triples
Learning	unsupervised	trained on AIDA	trained on a Wikipedia sample

Table 7: Comparison of entity linking pipeline architectures. VINCULUM components are described in detail in Section 4, and correspond to Figure 2. Components found to be most useful for VINCULUM are highlighted.

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC	Overall
CrossWikis	80.4	85.6	86.9	78.5	62.4	62.6	62.4	87.7	70.3	75.0
+FIGER	81.0	86.1	86.9	78.8	63.5	66.7	64.5	87.7	75.4	76.7
+Coref	80.9	86.2	87.0	78.6	59.9	68.9	66.3	87.7	86.6	78.0
+Coherence =VINCULUM	**81.4**	**86.8**	**87.0**	79.9	63.7	69.4	67.5	**88.5**	86.9	79.0
AIDA	73.2	78.6	77.5	68.4	52.0	71.9	**74.8**	77.8	75.4	72.2
WIKIFIER	79.7	86.2	86.3	**82.4**	**64.7**	**72.1**	69.8	85.1	**90.1**	**79.6**

Table 6: **End-to-end performance** (%): We compare VINCULUM in different stages with two state-of-the-art systems, AIDA and WIKIFIER. The column "Overall" lists the average performance of nine data sets for each approach. CrossWikis appears to be a strong baseline. VINCULUM is 0.6% shy from WIKIFIER, each winning in four data sets; AIDA tops both VINCULUM and WIKIFIER on AIDA-test.

Further reading

- *Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8). ACM.*

- *Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... & Weikum, G. (2011, July). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 782-792). Association for Computational Linguistics.*

- *Ling, X., Singh, S., & Weld, D. S. (2015). Design challenges for entity linking. Transactions of the Association for Computational Linguistics, 3, 315-328.*

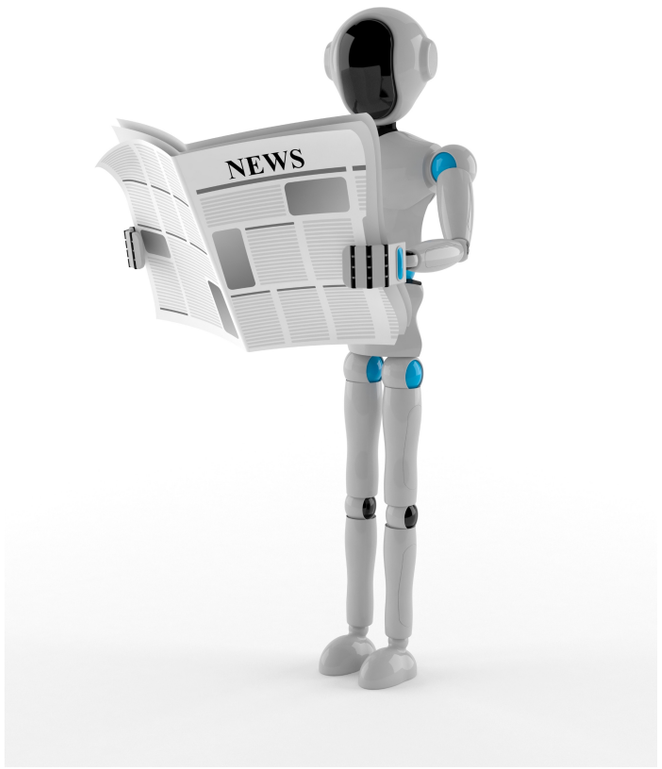
- *Van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J. (2016). Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. LREC*

FIGER (Ling & Weld 2012)

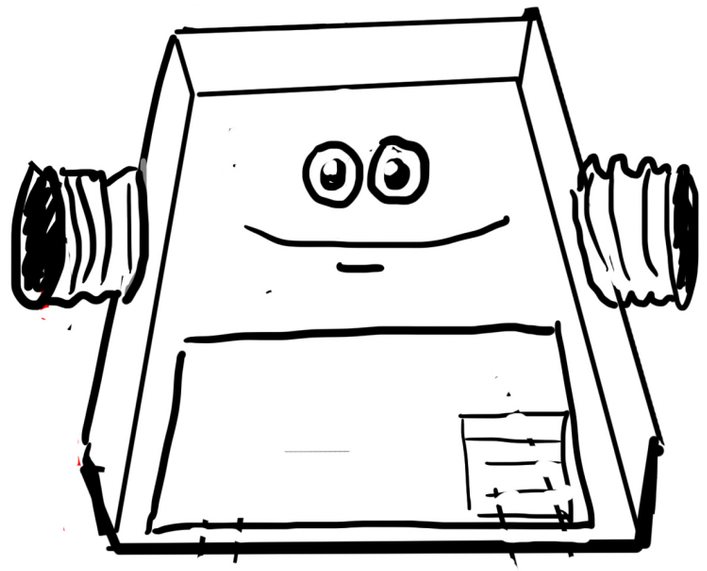
person	doctor	**organization**	terrorist_organization
actor	engineer	airline	government_agency
architect	monarch	company	government
artist	musician	educational_institution	political_party
athlete	politician	fraternity_sorority	educational_department
author	religious_leader	sports_league	military
coach	soldier	sports_team	news_agency
director	terrorist		
location	body_of_water	**product**	camera
city	island	engine	mobile_phone
country	mountain	airplane	computer
county	glacier	car	software
province	astral_body	ship	game
railway	cemetery	spacecraft	instrument
road	park	train	weapon
bridge			
art	written_work	**event**	military_conflict
film	newspaper	attack	natural_disaster
play	music	election	sports_event
		protest	terrorist_attack
building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

Figure 2: 112 tags used in FIGER. The bold-faced tag is a rough summary of each box. The box at the bottom right corner contains mixed tags that are hard to be categorized.

The case of NewsReader



ICT 316404, FP7-ICT-2011-8
www.newsreader-project.eu

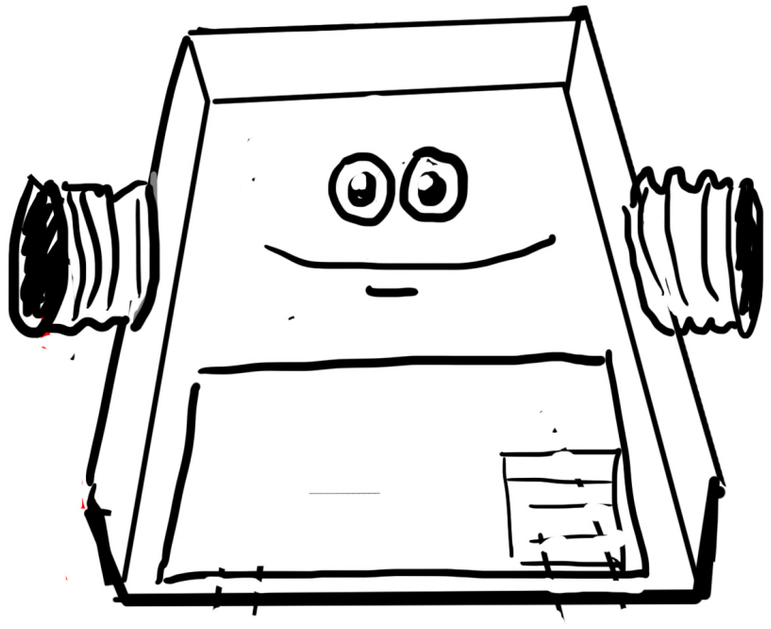


https://youtu.be/rYLaVN3oqLI

The case of NewsReader
a reading machine

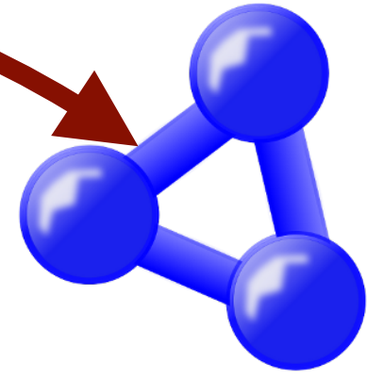
automotive industry
2003 - 2015

2.3 million articles
50 million entity mentions



Semantic Web
RDF-TRIPLES

what -who - where - when

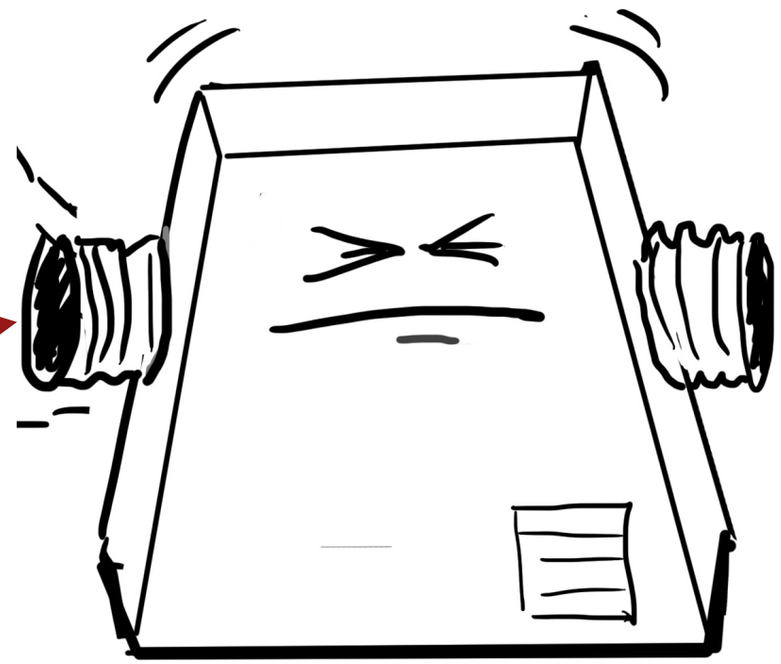


1.2 billion statements
2.2 million people, organisations, places

Ask instead of Read

Dark entities and false identities

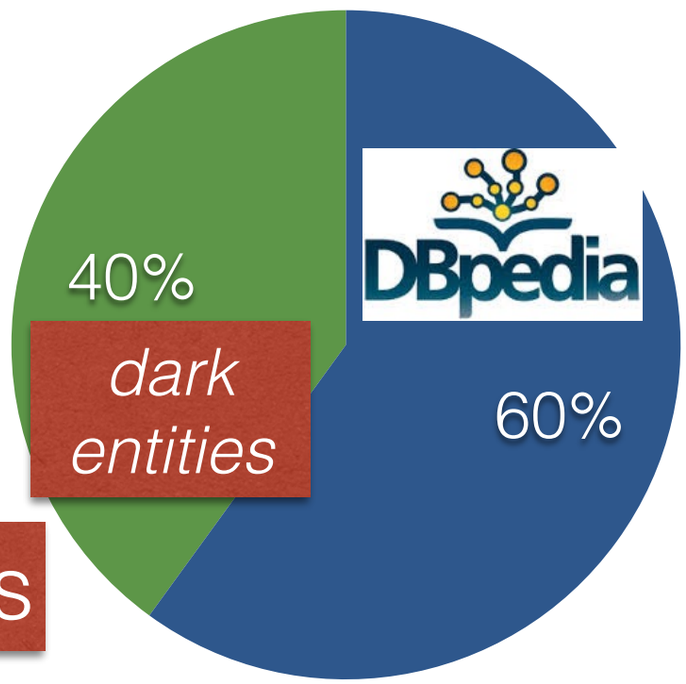
automotive industry
2003 - 2015



Semantic Web
RDF-TRIPLES

what -who - where - when

2.3 million articles
50 million entity mentions



1.2 billion statements
2.2 million people, organisations, places

40% ‘dark’ entities

60% ‘known’ entities

Entities

	in DBpedia	generated	Total	% DBpedia
linked Persons	121,834	530,257	652,091	18.68%
linked Organizations	59,032	814,088	873,120	6.76%
linked Locations	86,236	188,130	274,366	31.43%

Organizations	Number of Mentions
dbpedia:Toyota	130,783
dbpedia:General_Motors	77,420
dbpedia:Ford_Brasil	69,222
dbpedia:Republican_Party_(United_States)	53,106
dbpedia:Volvo	51,897
dbpedia:Democratic_Party_(United_States)	50,768
dbpedia:Volkswagen	48,748
dbpedia:Saab_Group	45,346
dbpedia:UEFA	40,692
dbpedia:Chrysler	39,572
dbpedia:Renault	39,443
dbpedia:UD_Trucks	37,719
dbpedia:Porsche	32,863
dbpedia:Daimler_AG	30,555
dbpedia:Honda	29,650
dbpedia:Toyota_Motorsport_GmbH	28,513
dbpedia:Opel	28,164
dbpedia:Volkswagen_Group	26,953
dbpedia:BMW	26,854
dbpedia:Suzuki	25,260
dbpedia:Ford_Germany	25,030
dbpedia:Fiat	24,280
dbpedia:Jaguar_Cars	23,415
dbpedia:Toyota_Manufacturing_UK	22,996
dbpedia:Proton_(automobile)	20,994
dbpedia:Mazda	20,438
dbpedia:Audi	20,330
dbpedia:Mercedes-Benz	19,510
dbpedia:Airbus	18,481

Persons	Number of Mentions
dbpedia:Abraham_Lincoln	68,904
dbpedia:Andrew_Johnson	26,306
dbpedia:Barack_Obama	19,761
dbpedia:Presidency_of_Barack_Obama	16,037
dbpedia:United_Nations	12,819
dbpedia:Zachary_Taylor	11,713
dbpedia:George_W._Bush	10,891
dbpedia:Gray_Davis	10,670
dbpedia:Ron_Paul	10,140
dbpedia:List_of_Heroes_characters	9,968
dbpedia:Gerry_Anderson	9,888
dbpedia:John_Howard	9,206
dbpedia:Walter_Scott	9,198
dbpedia:George_Washington	8,814
dbpedia:Ralf_Schumacher	8,765
dbpedia:Carlos_Ghosn	8,497
dbpedia:John_G._Robinson	7,866
dbpedia:Lewis_Hamilton	7,742
dbpedia:Trecia-Kaye_Smith	7,728
dbpedia:Fernando_Alonso	7,702
dbpedia:John_Adams	7,091
dbpedia:Robert_E._Lee	7,059
dbpedia:James_Baker	7,031
dbpedia:Celtic_art	6,868
dbpedia:Don_Murphy	6,433
dbpedia:David_Cameron	6,219
dbpedia:List_of_Sons_of_Anarchy_char	5,946
dbpedia:Lyle_Campbell	5,911
dbpedia:Jock_Young	5,749

Locations	Number of Mentions
dbpedia:England	105,342
dbpedia:Japan	102,530
dbpedia:United_States	98,872
dbpedia:Europe	97,154
dbpedia:Taiwan	86,450
dbpedia:India	84,941
dbpedia:Australia	72,135
dbpedia:Germany	56,223
dbpedia:Canada	51,164
dbpedia:Chinese_people	50,120
dbpedia:Americas	48,707
dbpedia:Russia	45,643
dbpedia:Empire_of_Japan	36,517
dbpedia:France	36,024
dbpedia:Billboard_200	30,834
dbpedia:London	28,553
dbpedia:China	27,342
dbpedia:Nebraska	27,141
dbpedia:Detroit	26,778
dbpedia:New_York	26,555
dbpedia:Second_Spanish_Republic	25,175
dbpedia:Spain	23,984
dbpedia:Malaysia	23,972
dbpedia:Tsardom_of_Russia	23,820
dbpedia:Italy	22,569
dbpedia:Chicago	21,880
dbpedia:Michigan	21,785
dbpedia:Australia_national_women's_cricket_team	21,617
dbpedia:North_America	21,566

Missing entities: NIL or Dark entities

- Dark Entities: domain entities may not be present

Ford Names Paul Bellew It's First Chief Data and Analytics Officer

By andreaharrison FOLLOW | 22 December 2014 | Like 0

SHARE ON FACEBOOK

Ford Motor Co.'s hiring last week of a new senior officer illustrates the auto industry's accelerating dive into big data.

Ford taps for General Motors exec to help it find its way in big data

Automotive News reports that the company named Paul Ballew, a former top sales analyst at General Motors, to the new post of chief data and analytics officer. Mr. Ballew, 50, will oversee the automaker's global analytics efforts

Missing entities: NIL or Dark entities

- Dark Entities: domain entities may not be present

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
Upload file
Special pages
Printable version

Languages

Not logged in Talk Contributions Create account Log in

Special page

paul ballew

Search results

Help

paul ballew

Search

Results 1 - 20 of 125

Content pages Multimedia Everything Advanced

Did you mean: paul ***bailey***

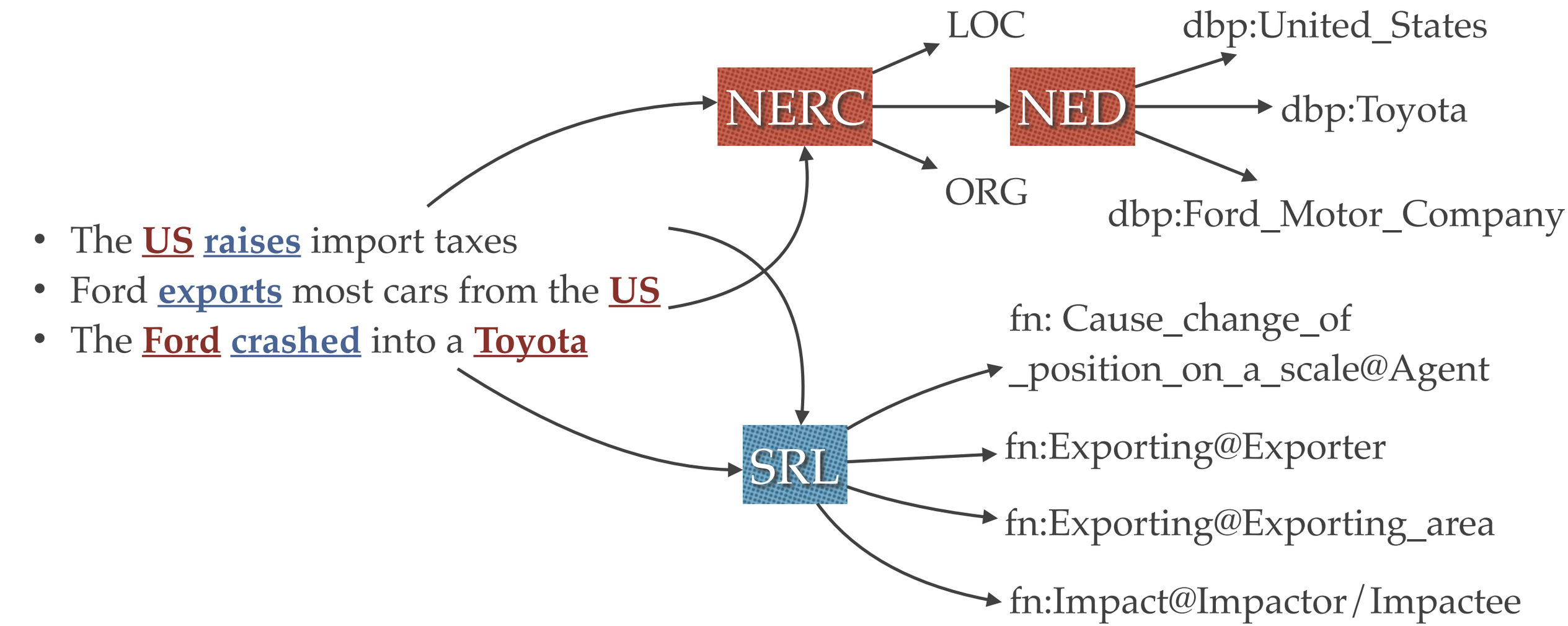
The page "Paul ballew" does not exist. You can ask for it to be created, but consider checking the search results below to see whether the topic is already covered.

Kurt Knoff (category People from the Minneapolis–Saint **Paul** metropolitan area)
post-football career as a commercial real estate broker in Minneapolis – Saint **Paul**
Ballew, Bill. Tough Enough to be Vikings: Minnesota's Purple Pride from A to
5 KB (503 words) - 15:10, 12 January 2016

Billy **Ballew** Motorsports
Billy **Ballew** Motorsports is a team that competes in the NASCAR Camping World
Truck Series. They were formed in 1996 by Georgia businessman Billy **Ballew**.

Identity crisis

- The **US** **raises** import taxes
- Ford **exports** most cars from the **US**
- The **Ford** **crashed** into a **Toyota**



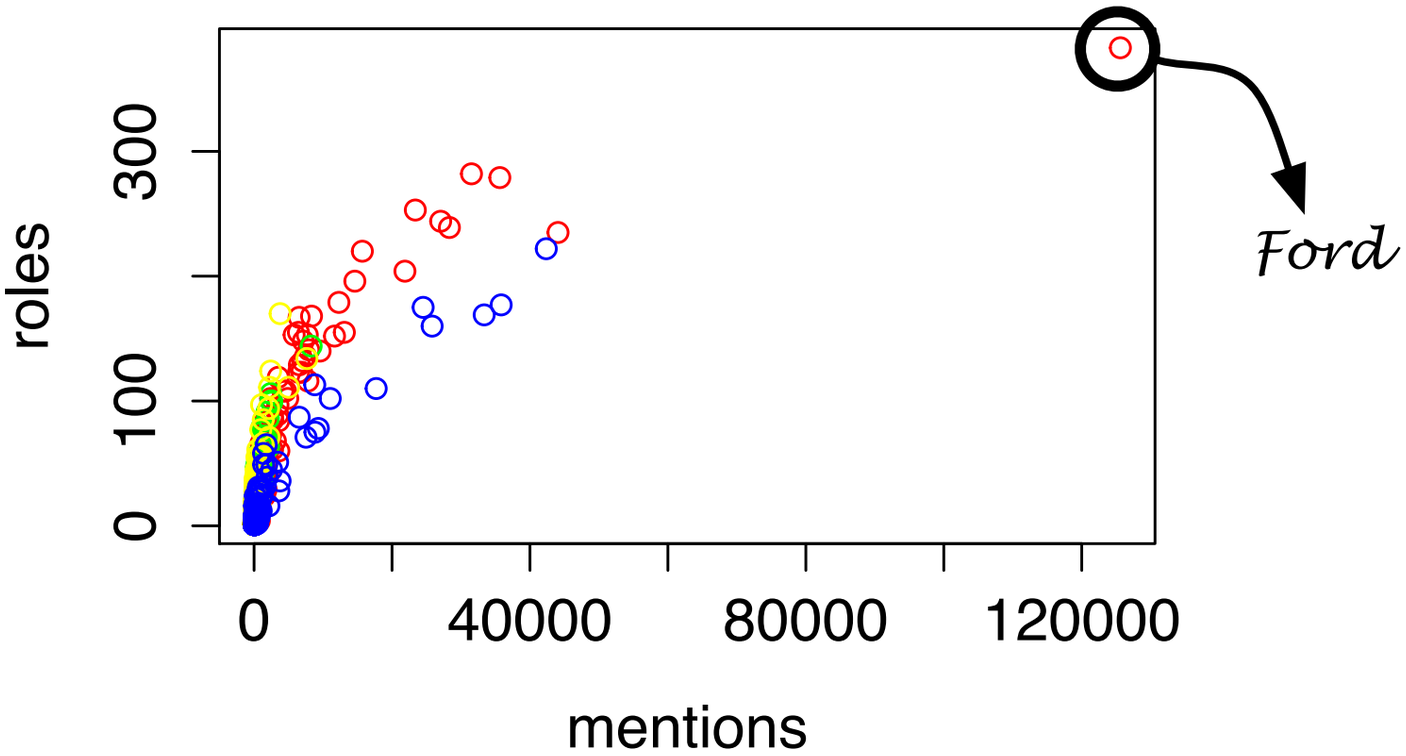
Identity crisis

Table 2. Occurrences as Actor or Places per Class

Class	Actor (total)	Place (total)	Actor (%)	Standard Deviation
Country	152,708	132,341	53,57%	40.93%
Person	106,318	9,447	91.84%	21.56%
Company	609,971	84,378	87.85%	29.63%
Motor Company	431,619	63,917	87,10%	28.46%
Automobile	116,893	5,804	95.27%	14.45%

Based on 63K news articles on the automotive industry

383 Roles of Ford



5 Experiencer
5 Goods
5 Message
5 Protagonist
6 Entity
6 Event
7 Addressee
7 Item
7 Phenomenon
8 Communicator
8 Goal
8 Source
9 Topic
12 Cognizer
13 Speaker
18 Cause
20 Theme
40 Agent

4 Cause_harm
4 Cause_to_move_in_place
4 Collaboration
4 Emptying
4 Grant_permission
4 Operate_vehicle
4 Partiality
4 Purpose
4 Questioning
4 Removing
4 Request
4 Ride_vehicle
4 Self_motion
4 Sending
4 Statement
4 Telling
5 Change_of_leadership

Entities resolved?

- 36% dark entities in news
- Many disambiguation errors

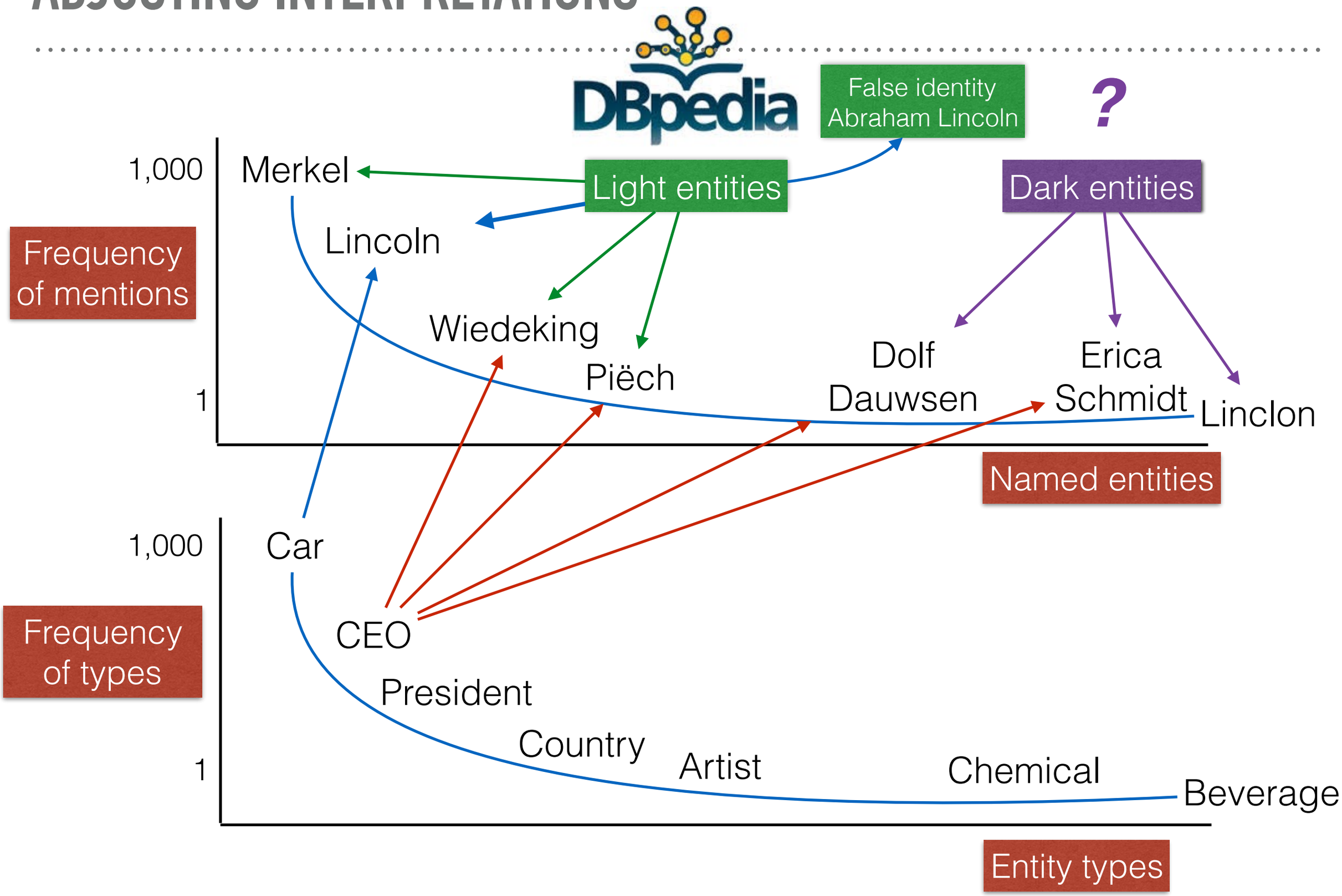
dominant domain resolver:
Peter
Schreyer
Peter Schreyer

- Many dark entities linked to light entities
- Within the same document
- Within the same domain

relevant time resolver
Abraham Lincoln's
relevance for 2003-2015

- Why link Lincoln to the president in the context of automotive industry?

ADJUSTING INTERPRETATIONS



Learning about the world

60%

DBpedia

Dark

DBpedia

40%

dbp:OfficeHolder

Mark Fields

darkdbp:DataAnalyticsOfficer

Paul Ballew

Further reading

- *Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8). ACM.*

- *Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... & Weikum, G. (2011, July). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 782-792). Association for Computational Linguistics.*

- *Ling, X., Singh, S., & Weld, D. S. (2015). Design challenges for entity linking. Transactions of the Association for Computational Linguistics, 3, 315-328.*

- *Van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J. (2016). Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. LREC*