University of Stuttgart
Institute for Natural Language Processing

Collaborations with
Abhijeet Gupta[1]
Marco Baroni[2]
Gemma Boleda[2]
Gabriella Lapesa[1]
V Thejas[3]
Matthijs Westera[2]

[1] University of Stuttgart
[2] UPF Barcelona
[3] BITS Pilani

Entities as a Window into (Distributional) Semantics

Sebastian Padó


Macron tells PM revised Brexit deal 'not an option'
The leaders will meet hours after the French president insisted the Brexit deal could not be reopened.
UK POLITICS

Merriam-Webster

deal noun (2)

Definition of *deal* (Entry 3 of 4)

1 *business*

a : an act of dealing (see DEAL entry 2 sense 3)

// big corporate *deals*

// a real estate *deal*

b : BARGAIN

// got a great *deal* on a new TV

// accepted a plea *deal*

c : CONTRACT sense 1a

// signed a 2-year *deal*

option noun

op·tion | \ ˈäp-shən \

Definition of *option* (Entry 1 of 2)

1 : an act of choosing

// hard to make an *option* between such alternatives

2 **a** : the power or right to choose : freedom of choice

// He has the *option* to cancel the deal.

b : a privilege of demanding fulfillment of a contract on any day within a specified time

c : a contract conveying a right to buy or sell designated securities, commodities, or property interest at a specified price during a stipulat period

also : the right conveyed by an option

// The ad is for a condo to rent with an *option* to buy.

d : a right of an insured person to choose the form in which payments du policy shall be made or applied

3 : something that may be chosen: such as

a : an alternative course of action

// didn't have many *options* open

- *deal, option* are **categories (concepts)**
  - Listed in dictionary

- *Macron, Brexit* are individual **entities/events**
  - Listed in encyclopedia

Emmanuel Macron

Macron in 2017

President of France

Incumbent

Assumed office
14 May 2017

Prime Minister Édouard Philippe

Preceded by François Hollande

Brexit

Articles on the withdrawal of the United Kingdom from the European Union.

Main topics

Brexit negotiations (2017 · 2018 · 2019) · No-deal Brexit · Impact of Brexit on the European Union · Brexit and arrangements for science and technology · Economic effects of Brexit · Opposition to Brexit in the United Kingdom

Secondary topics

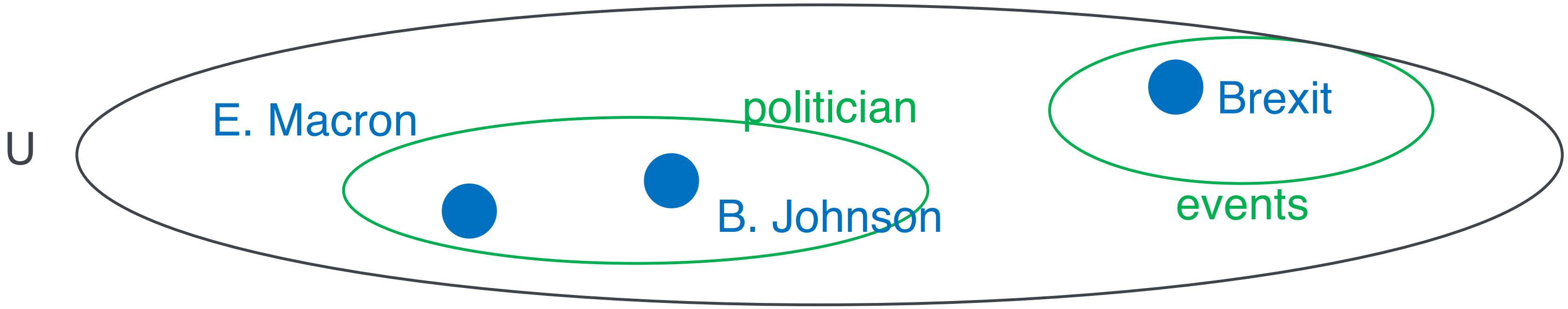
Withdrawal from the European Union · Euroscepticism in the United Kingdom

Referendum

European Union Referendum Act 2015 · Campaigning in the United Kingdom European Union membership referendum · European Union Act 2016 (Gibraltar) · Issues in the United Kingdom European Union membership referendum · Opinion polling for the United Kingdom European Union membership referendum · Endorsements in the United Kingdom European Union membership referendum · International reactions to the 2016 United Kingdom European Union membership referendum

Model-theoretic semantics

- Meaning of language units defined relative to world model (Gamut 1991: Universe U = set of individuals)
- Proper nouns and other entities:
  - Mapped onto elements of the universe
- Common nouns, adjectives, and other categories:
  - Mapped onto sets of elements of the universe

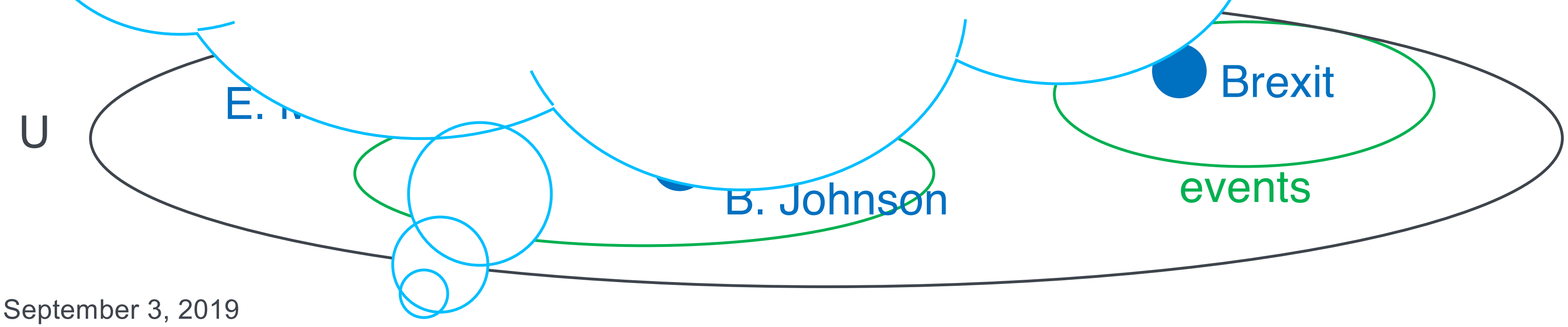


Model-theoretic semantics

- Meaning of langua… its … ve to world model (… lividuals)
- Prope…
  - M…
- Con… es:
  - M… iverse

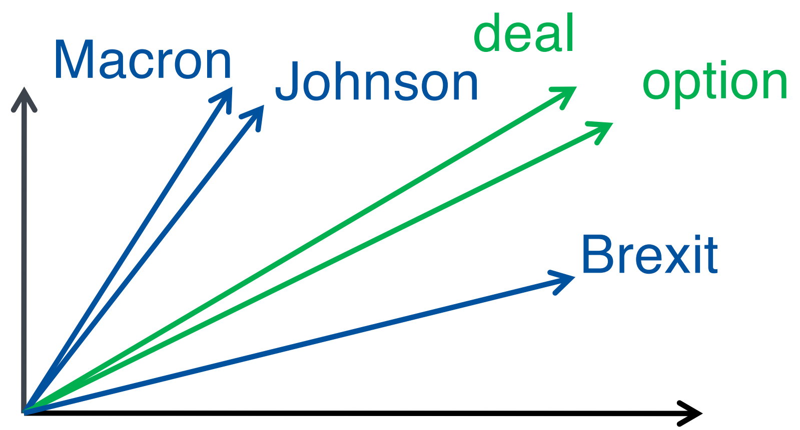
Entities and categories
are fundamentally different

What about current NLP?



Distributional Semantics (DS)

- Dominant paradigm to acquire lexical information:
  - Learn linear algebra representations of linguistic units from **context**
  - A.k.a. Vector spaces, embeddings, distributed representations



- Still DS because all use the **“distributional hypothesis”**: “You shall know a word by the company it keeps” (Firth, Harris, Miller & Charles 1991, etc.)

Distributional Semantics (DS)

- Dominant paradigm … mation:
  - Learn li… represe… uni…
  - A.k… emb…
- Still DS … hypothesis": "You shall kn… ompany it keeps" (Firth, Harris, Mil… Charles 1991, etc.)

deal

option

Brexit

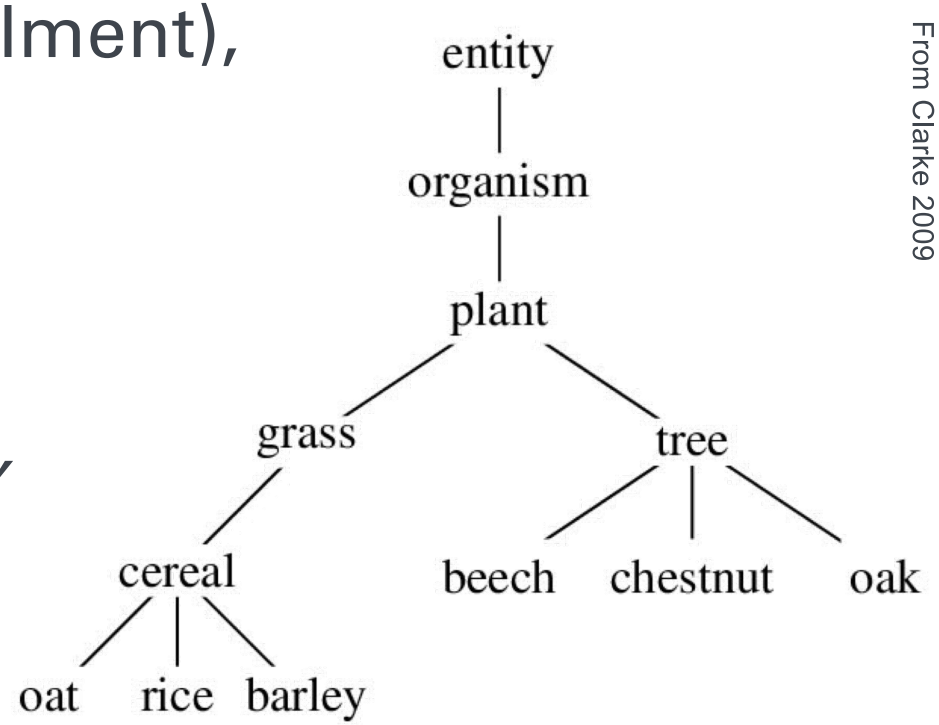
How is this applied to categories / entities in NLP?

Split by subcommunity

Computational Lexical Semantics

- Strong focus on modelling linguistic aspects of meaning: categories and relations among categories
  - Hyponymy/hypernymy (entailment), synonymy, meronymy
  - Also diachronic change

"Interested in generalizations"

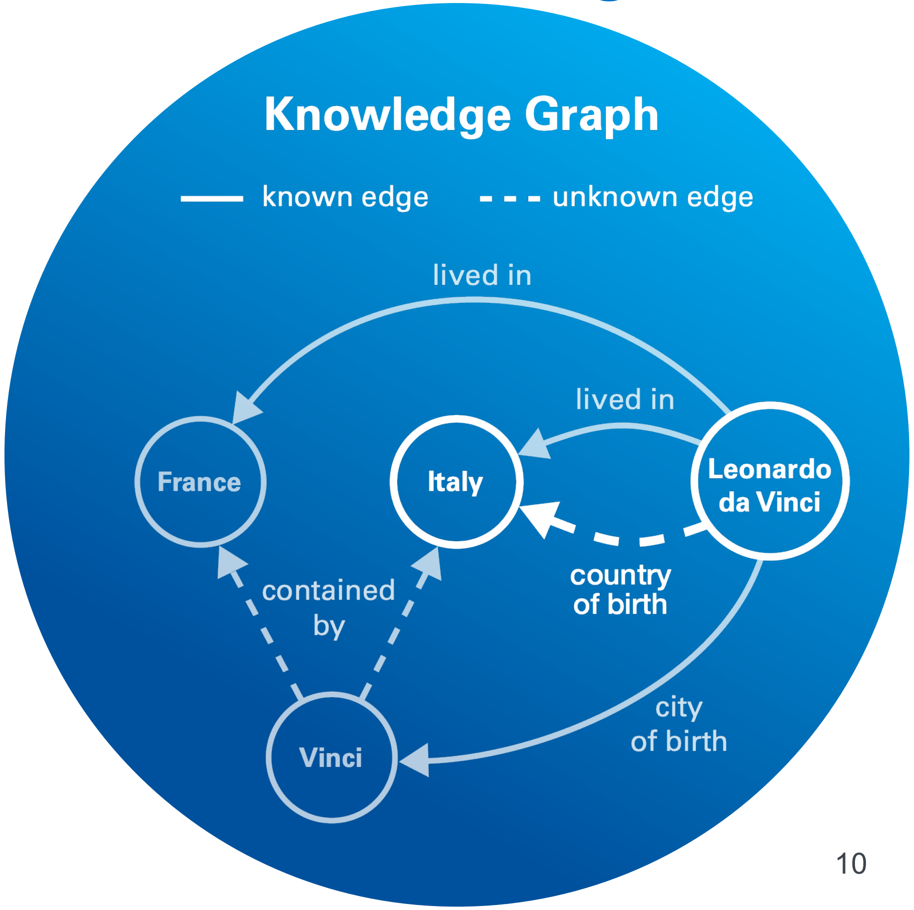


From Clarke 2009

Semantic Web / Information Extraction

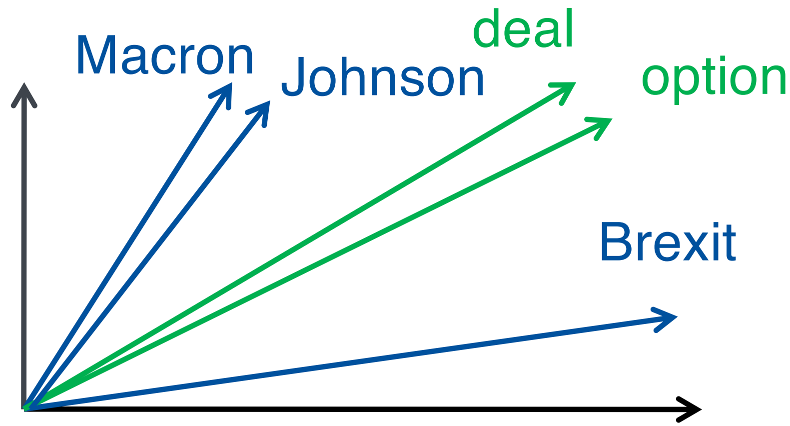
- Complementary focus on modelling world knowledge aspects of meaning: entities and relations among entities
- Knowledge bases / knowledge graphs

"Interested in particularities"



The Current Situation

- So Distributional Semantics is applied
  - to both entities and categories
  - to learn fairly different things
  - How is this possible?



- **“It just works”**
  - **DS is a practice without a theory**

Agenda for this presentation

- Q: Are there relevant differences in the way we can apply DS to modelling entities and categories?
- Research strand 1: Knowledge Bases
  - How far can we push DS in learning world knowledge?
- Research strand 2: The Instantiation Relation
  - How do categories and entities behave distributionally?

Benefit: insights into **capabilities** and **limits** of distributional approaches to meaning

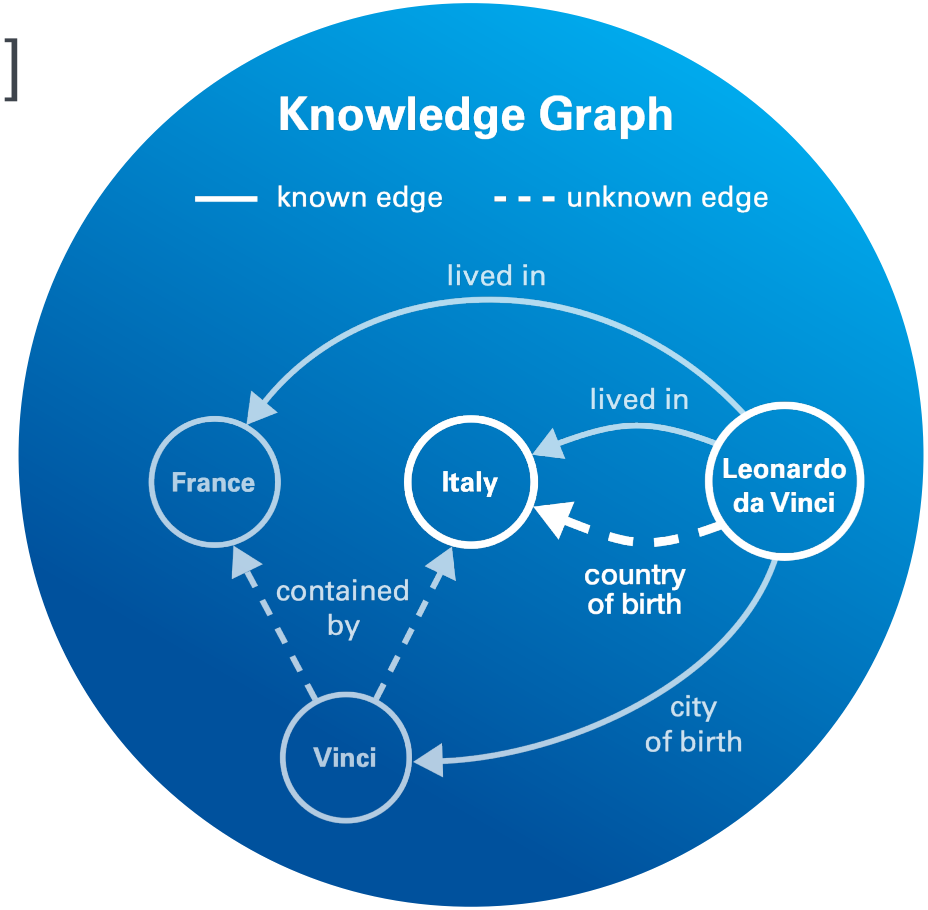
Agenda for this presentation

- Q: Are there relevant differences in the way we can apply DS to modelling entities and categories?
- **Research strand 1: Knowledge Bases**
  - **How far can we push DS in learning world knowledge?**
- Research strand 2: The Instantiation Relation
  - How do categories and entities behave distributionally?

Benefit: insights into **capabilities** and **limits** of distributional approaches to meaning

Strand 1: Knowledge Base Completion

- Challenge: KBs are incomplete [Min et al. 2013, West et al. 2014]
  - **Knowledge Base Completion (KBC)**: Add missing edges to knowledge graph
- Very active area of research
  - Representation learning
  - Learn embeddings for entities and relations



Entity Embeddings and KBC

- KBC embeddings can be learned from text, KB, or both
  - Our Interest: limits of distributional semantics
    - Focus on text-based embeddings of entities
- Entities have fine-grained attributes with specific values
- **Research Question: Can all attributes be predicted from vanilla word embeddings? (And if not, why not?)**

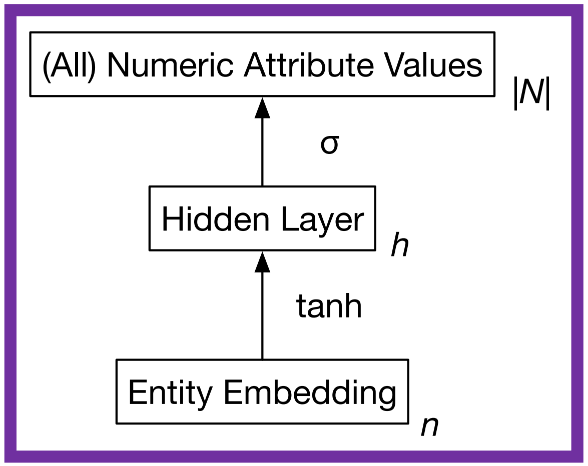
$$\textbf{Italy}\left\{\begin{array}{ll} \text{sunny} & 30 \\ \text{wine} & 15 \\ \text{beach} & 12 \\ \text{Rome} & 10 \\ \text{Naples} & 6 \end{array}\right\} \longrightarrow \textbf{Italy}\left\{\begin{array}{ll} \text{Population} & \text{: 61 million} \\ \text{Area} & \text{: 301,000 sq.km} \\ \text{Language} & \text{: Italian} \\ \text{Contained by} & \text{: Europe} \\ \text{Currency used} & \text{: Euro} \end{array}\right\}$$

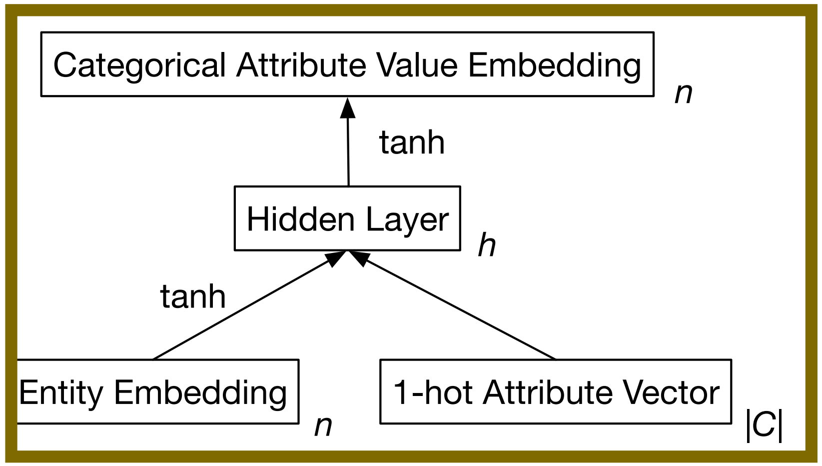
Simple Supervised KBC [Gupta et al. 15,17]

- Task: Use entity embeddings to predict entity attributes with Multi-Layer Perceptron (MLP)
  - Numeric: predict value(s)
  - Categorical: predict embedding for relatum (Italy, currency, Euro)

Italy

Population : 61 million
Area : 301,000 sq.km
Language : Italian
Contained by : Europe
Currency used: Euro





Evaluation of Attributes

- Categorical attributes: Mean Reciprocal Rank (MRR)
  - Mean rank of predicted relatum embedding among nearest neighbors of true relatum embedding

- Numeric attributes: Correlation
  - Spearman correlation between predicted and true rankings of entities w.r.t. attribute

(Leaving out details here; see papers)

Experimental Setup

- **Embeddings:** Google News vectors (Mikolov et al. 2013)
  - Word2Vec skipgram, 300 dimensions
- **Experimental setup:** Train/Test on 7 FreeBase domains

Domain	# Entities (train/val/test)	$\|C\|$	$\|N\|$
Animal	279/93/93	22	118
Book	16/5/6	8	2
Citytown	1783/594/595	57	62
Country	155/53/51	79	698
Employer	720/140/141	50	55
Organization	187/63/62	36	32
People	85/28/29	25	76
Sum	3225/976/977	277	1043

Experimental Setup

- **Embeddings:** Google News vectors (Mikolov et al. 2013)
  - Word2Vec ...
- **Experim**... s

Three case studies / observations

(My) explanation to follow

			698
			55
		36	32
Peo		25	76
Su	3225/976/977	277	1043

Domain Country: Numeric Attributes

Feature	Correlation of MLP
Geolocation (Lat. / Long.)	**0.93**
GDP_per_capita	**0.89**
CO2_emissions_per_capita	**0.88**
…	…
GDP_nominal	**0.78**
…	…
Date_founded	0.54
Religion_percentage	0.42

best

worst

- Attributes differ greatly in difficulty
  - Geographical attributes easy (Louwerse et al. 2009)

Geolocation: The Good

Actual

Predicted

A	Hong Kong
B	Bangladesh
C	Cocos Islands
D	Eritrea
E	Latvia
F	Belarus
G	Iran

Geolocation: The Bad

A	New Caledonia
B	Cocos Islands
C	Cook Islands
D	Mauritius

E	Niue
F	Tuvalu
G	Vanuatu

Domain Country: GDP

Feature	Correlation of MLP
Geolocation (Lat. / Long.)	**0.93**
GDP_per_capita	**0.89**
CO2_emissions_per_capita	**0.88**
…	…
GDP_nominal	**0.78**
…	…
Date_founded	0.54
Religion_percentage	0.42

best

worst

- Even very similar attributes differ substantially (?)

Domain Country: Difficult Attributes

Feature	Correlation of MLP
Geolocation (Lat. / Long.)	**0.93**
GDP_per_capita	**0.89**
CO2_emissions_per_capita	**0.88**
...	...
GDP_nominal	**0.78**
...	...
Date_founded	0.54
Religion_percentage	0.42

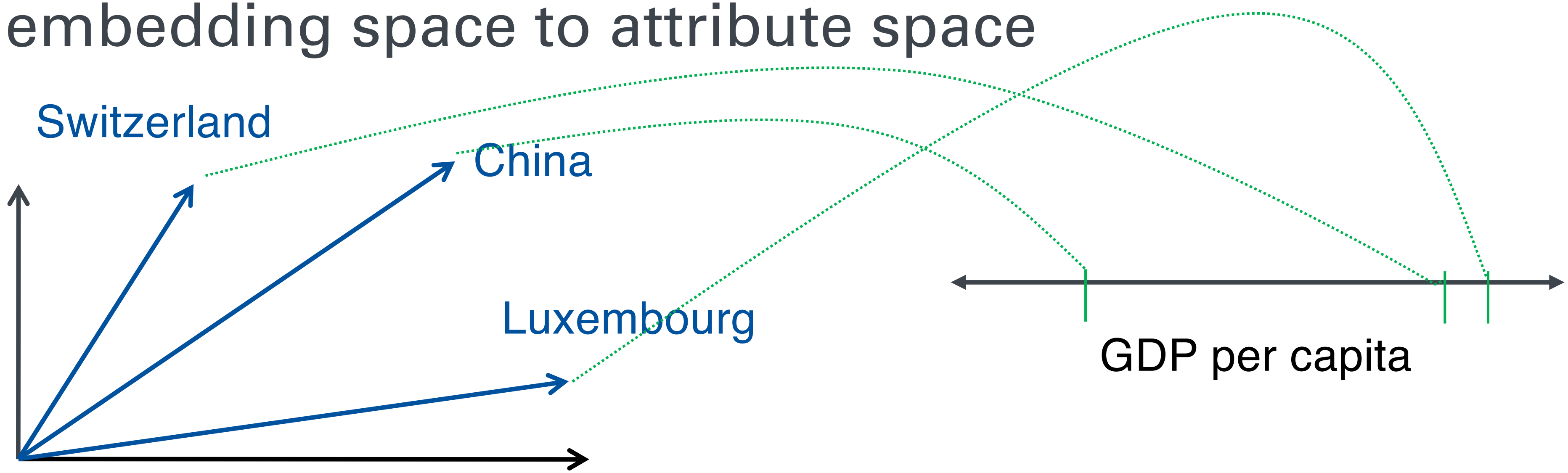
best

worst

- The most difficult attributes appear to be very **specific**

Contextual Support

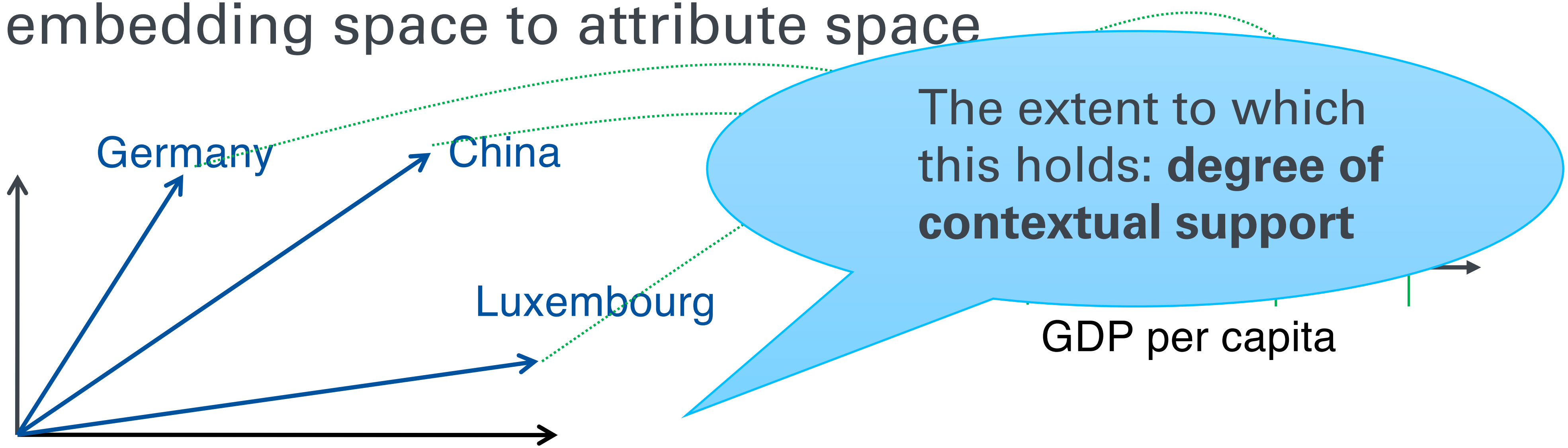
- Our KBC task = learn mappings from **context-derived** embedding space to attribute space



1. Attribute must correlate with **prominent context cues**
2. Entities with **similar values** of attribute must co-occur with **similar context cues**

Contextual Support

- Our KBC task = learn mappings from (BOW) embedding space to attribute space



1. Attribute must correlate with **prominent context cues**
2. Entities with **similar values** of attribute must co-occur with **similar context cues**

Contextual Support Accounts for…

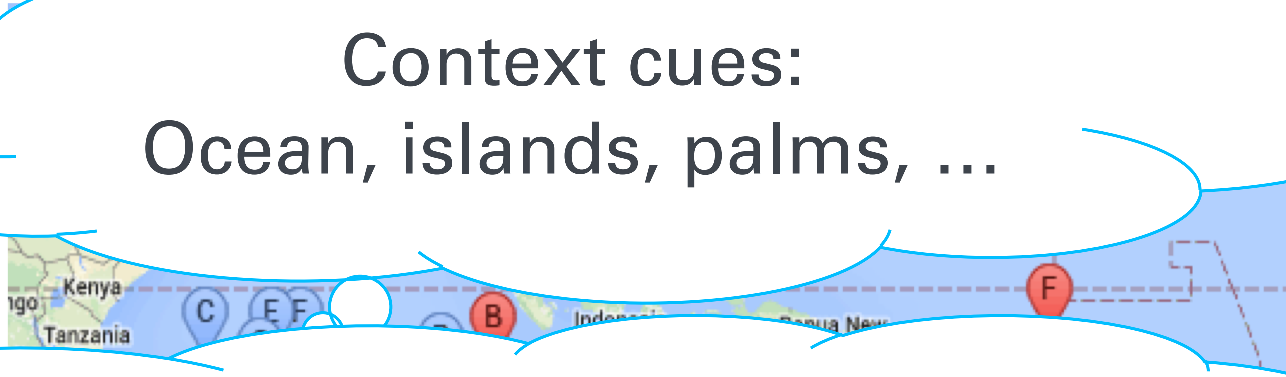
- The island displacement: “Hubness effect”
  - Predictions for sparse entities dominated by similar, more frequent entities

A	New Caledonia
B	Cocos Islands
C	Cook Islands
D	Mauritius

E	Niue
F	Tuvalu
G	Vanuatu

Contextual Support Accounts for…

- The island displacement: "Hubness effect"
  - Predictions for sparse entities dominated by similar frequent entities



Context cues:
Ocean, islands, palms, ...

Hubs with these contexts:
Maldives, Seychelles, Mauritius

C	Cook Islands
D	Mauritius

	Tuvalu
G	Vanuatu

Contextual Support Accounts for

- GDP_per_capita being easier than GDP_nominal
  - GDP per capita comes with **more consistent context cues**

GDP per capita

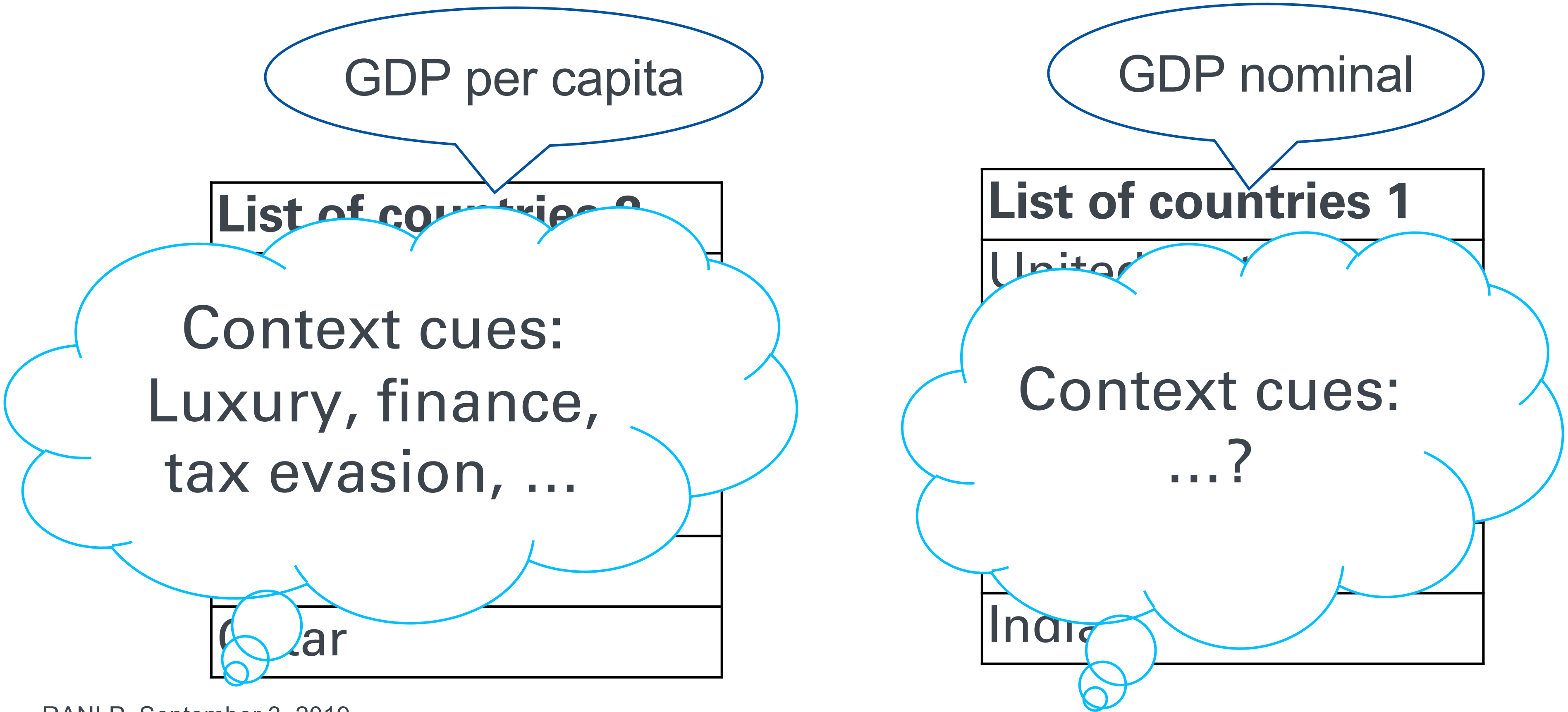
List of countries 2
Luxembourg
Switzerland
Norway
Ireland
Iceland
Qatar

GDP nominal

List of countries 1
United States
China
Japan
Germany
UK
India

Contextual Support Accounts for

- GDP_per_capita being easier than GDP_nominal
  - GDP per capita comes **with more consistent context cues**



Contextual Support Accounts for

- Difficulty of learning very specific attributes (date of foundation, countries exported to..)
  - Indicated by highly specific, low frequency context cues
  - "Drowned out" by other information in pretrained BOW vectors
  - Compare to pattern-based approach (Hearst 1992):

The modern state of Italy was created in the year 1861.
In 1861, Italy was largely unified.
The Kingdom of Italy was founded on this day in 1861.

→

Italy
Date founded : 1861
Area : 301,000 sq.km
Language : Italian
Contained by : Europe
Currency used: Euro

Take-home from Strand 1

- Knowledge can only be learned distributionally if has a substantial degree of **contextual support**
- Future directions:
  - Measuring / quantifying contextual support
  - Increasing contextual support
    - Fine-tuning on labeled data – not a panacea (?)
    - Present specific patterns to learner (Roller & Erk 2016)
    - Use meta-information about attribute-attribute relations

GDP_per_capita =
GDP_nominal / population

Agenda for this presentation

- Q: Are there relevant differences in the way we can apply DS to modelling entities and categories?
- Research strand 1: Knowledge Bases
  - How far can we push DS in learning world knowledge?
- **Research strand 2: The Instantiation Relation**
  - **How do categories and entities behave distributionally?**

Benefit: insights into **capabilities** and **limits**
of distributional approaches to meaning

Strand 2: Instantiation

[Gupta et al. EACL 2017, ArXiv]

- We introduce a new semantic relation: **instantiation**
  - Hypernymy -- relation between **two categories**

    [Baroni et al. 12, Roller et al. 14, Santus et al. 14, Levy et al. 15, etc.]
  - Instantiation -- relation between **entity and category**
    - Many-to-many, not reflexive, not symmetrical, not transitive

An Instantiation Dataset

- 22k pairs: 5.5k positive pairs + 3* 5.5k negative pairs
  - Positive: Group entity with category
    - "Instance hypernym" relation from WordNet
  - Negative 1: INVERSE (switch entity and category)
  - Negative 2: INST2INST(entity + random other entity)
  - Negative 3: NOTINST (entity + wrong related category)

Positive	Abraham Lincoln – POTUS	Mumbai – city
Inverse	POTUS – Abraham Lincoln	city – Mumbai
Inst2Inst	Abraham Lincoln – Duncan Grant	Mumbai – Vicksburg
NotInst-inClass	Abraham Lincoln – doctor	Mumbai – residential area

Modeling Instantiation

- Architecture: let's use an MLP again (1 hidden layer)
  - Inspiration: hypernymy classifier (Roller et al. 14)
  - Input: Embeddings for two words, e.g. $v = w_1||w_2$
  - Output: Binary decision (instantiation or not)

(We experimented with different variations)

Experimental Setup

- Own dataset
  - Train-dev-test split with memorization filtering
  - No entity or concept appears in more than one section

- Embeddings: Google News

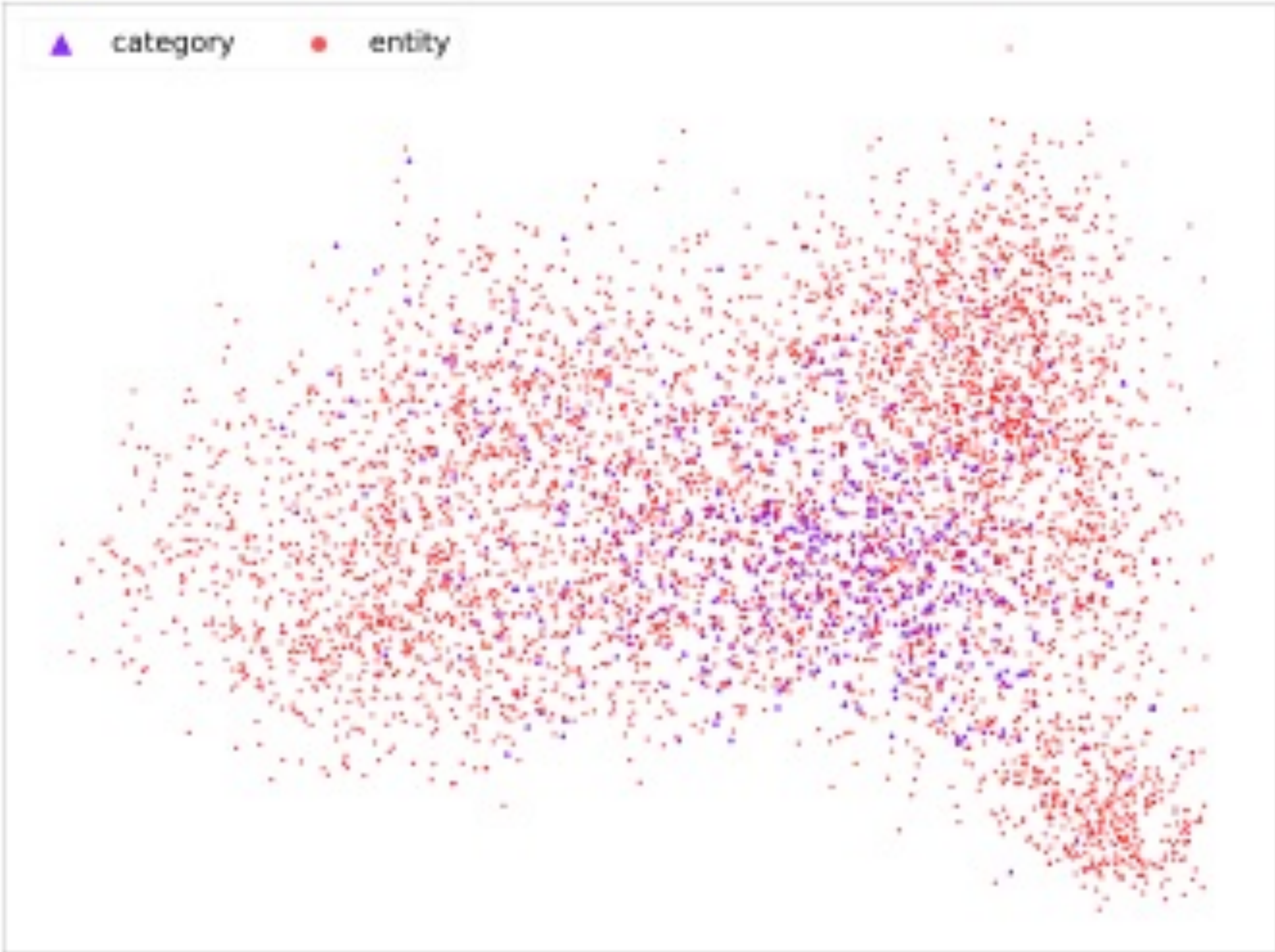
- Baseline: Always predict instantiation
- Evaluation: F1 for class instantiation

Results

Dataset	Pos:Neg	Neg ex.	BL_{pos}	MLP
Pos + *Inverse*	1:1	POTUS – Lincoln	0.67	0.96
Pos + *Inst2Inst*	1:1	Lincoln – Grant	0.67	0.91
Pos + *NotInst*	1:1	Lincoln – doctor	0.67	0.69
Pos + *Union*	1:3	*all*	0.40	0.63

- Inverse and Inst2Inst are very simple
  - Entities and categories are simple to distinguish
- NotInst is very difficult: hardly beats baseline!
  - Corresponds well to findings about hypernymy

What makes NotInst hard?



Category representation

- Our assumption: Embedding of **noun x** is a good representation of **category x**
  - (Universally assumed in lexical semantic modeling)
- That is actually questionable:
  - Informativity
  - Ambiguity, including metaphors
  - Lexical choice and speaker intent (Lapesa et al. 2017, Westera and Boleda 2019)

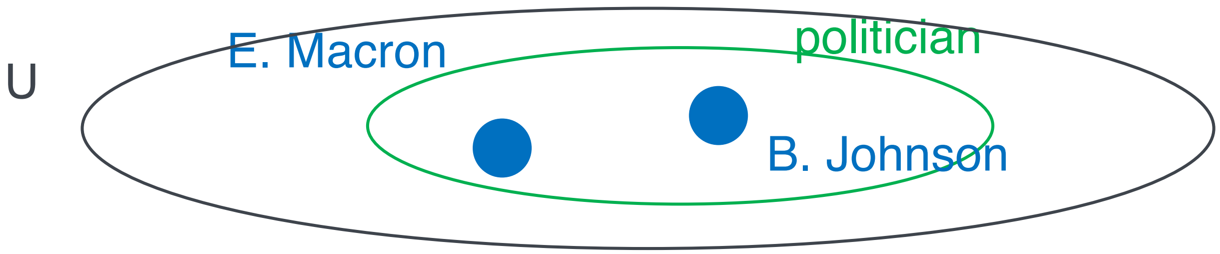
Grass is green

Elephant in the room

Fotograf vs. Fotografin (generic/female photographer)

Re-representing Categories

- Are there alternatives for concept representation?

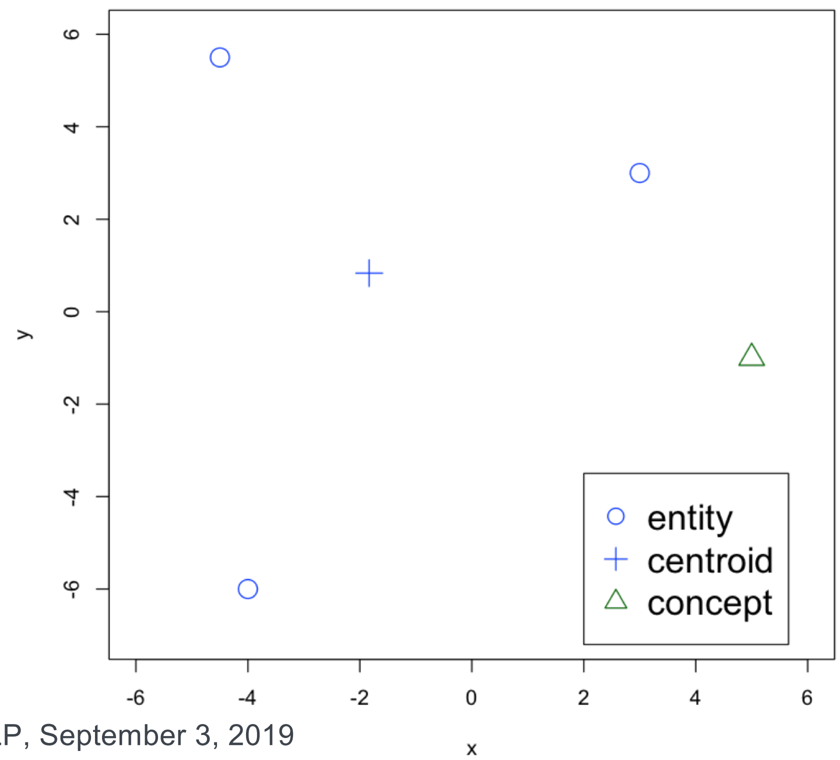


- Formal semantics: (extension of) concept = set of entities instantiating it
- In our context: Represent categories by the **centroid of their entity embeddings** („centroid embedding“)
  - Vs. traditional approach: „concept embedding“

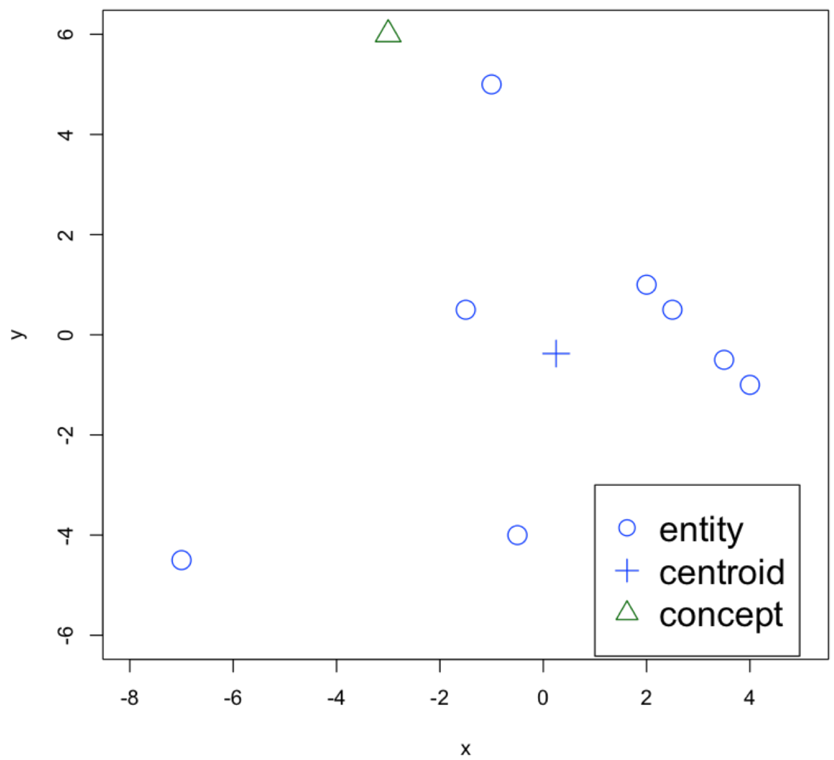
Does It Work?

$$\begin{array}{ll} \text{Entity} - \text{Entity} & \text{mean cos} = 0.22 \\ \text{Entity} - \text{Concept} & \text{mean cos} = 0.16 \\ \text{Entity} - \text{Centroid} & \text{mean cos} = 0.55 \end{array}$$

patriarch



geneticist



Experimental Validation

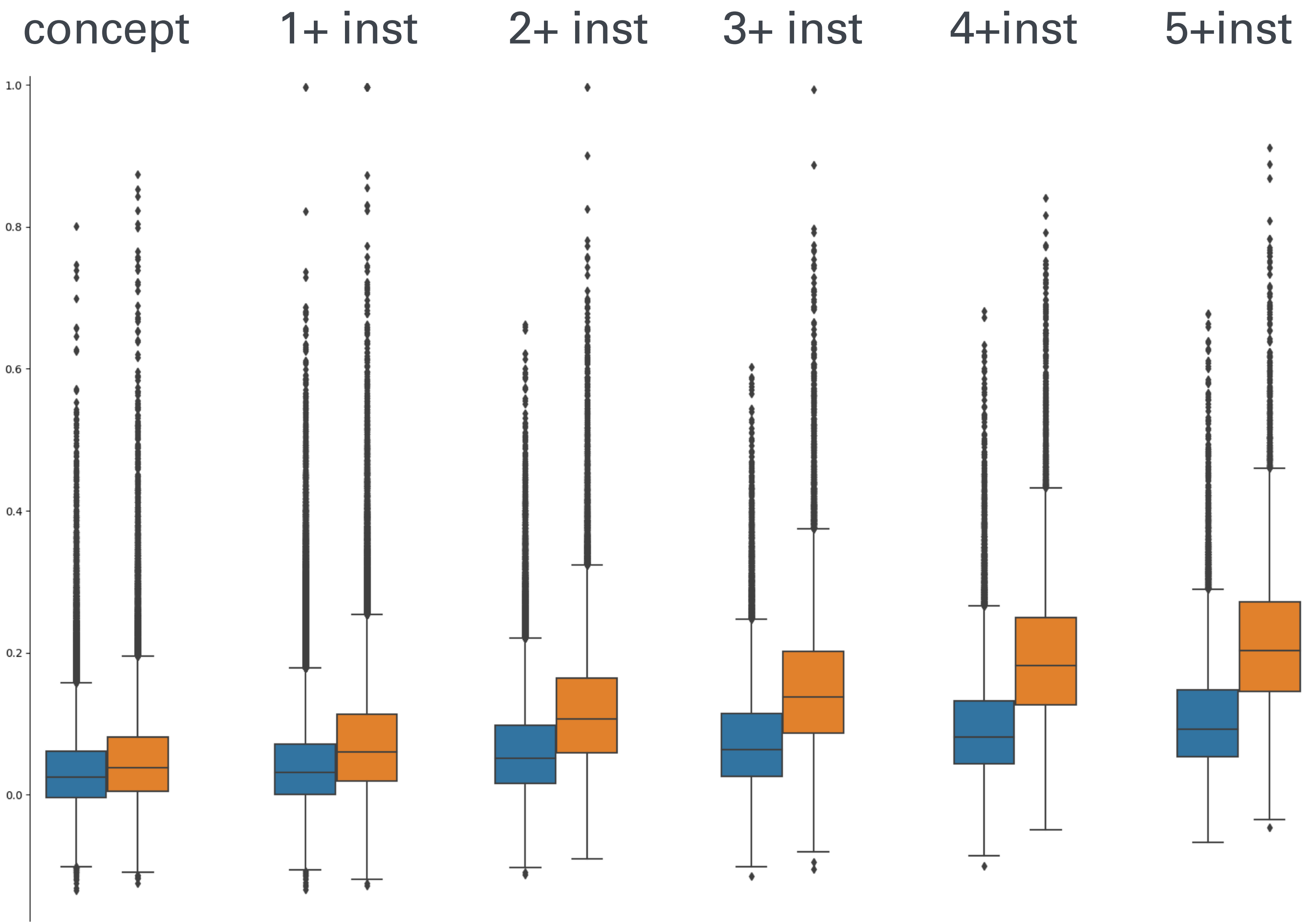
Dataset	Pos:Neg	BL_{Pos}	Concept emb.	Centroid emb.
Pos + *Inverse*	1:1	0.67	0.96	0.98
Pos + *Inst2Inst*	1:1	0.67	0.91	0.91
Pos + *NotInst*	1:1	0.67	0.67	**0.79**
Pos + *Union*	1:3	0.40	0.63	**0.76**

- Extension of previous experiment
  - Centroids built on training set
- **Improvement for centroid-based category representation**

Take-home from Strand 2

- Categories and entities **differ in distributional behavior**
  - Can be distinguished easily
  - But capturing entity-category relations is tricky
    - Analogy to difficult attributes in Strand 1
- How to improve comparability?
  - Here: Centroid-based representatio
    - Conceptually appealing
    - Requires more information about categories than just their names, namely instances

> How many instances are needed?
> Sneak preview!


concept
1+ inst
2+ inst
3+ inst
4+inst
5+inst
1.0
0.8
0.6
0.4
0.2
0.0

Wrap-up

- The Distributional Hypothesis – **usage determines meaning** – is at the heart of many NLP applications
  - But is it really true?
- My proposal today: Let’s relate the **properties of information we want to learn** to the **properties of the linguistic material** we want to learn it from
  - This presentation: Entities vs. categories
  - Other direction: Speaker intention vs. linguistic usage

University of Stuttgart

Thank you!

Sebastian Padó

e-mail pado@ims.uni-stuttgart.de
phone +49 (0) 711 685-81394
www.ims.uni-stuttgart.de/~pado

University of Stuttgart