University of Stuttgart

Experiments with category representations

Sebastian Padó

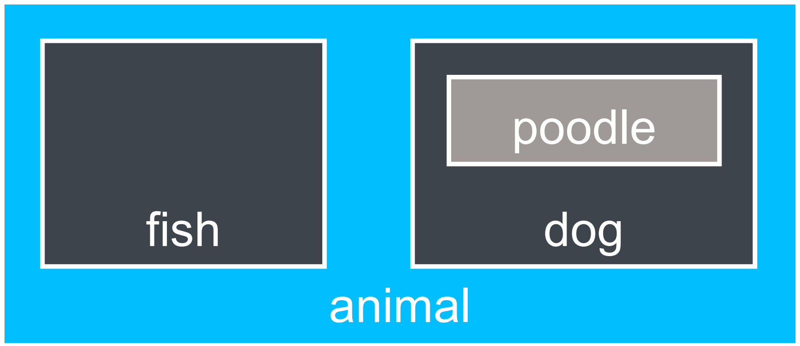
Collaborators in this work

- Abhijeet Gupta
- Jen Sikos
- Gemma Boleda
- Matthijs Westera

- Katrin Erk

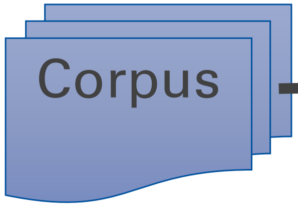
Categories

- “Categorization is the process of **forming categories** and **assigning objects** to them” (Murphy 2002)
  - Fundamental to our perception, understanding of the world
  - Categories are crucial by allowing **inferences**
    - Subsumption, Properties, Elementhood, …
- Arguably, crucial role in language



Relevance for computational linguistics

- Distributional semantics (Harris 1954; Miller & Charles 1991): Represent a word in terms of its occurrence contexts

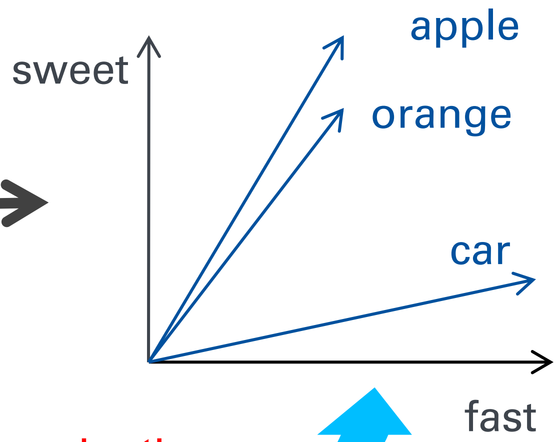


orange sweet

sweet **orange**

it went bad fast

	sweet	fast
orange	5	1
apple	4	1
car	1	20



Categorization: instances → senses

Categorization: senses → semantic classes

- Conceptual successor: Word **embeddings**

This presentation

- Use of word embeddings is directly related to categorization
  - Hypothesis: We can gain insights by examining word embeddings through the lens of categorization theory
- Two studies:
  1. How do current embedding models relate to categorization theories?
  2. How to properly learn categories from text, and the role of category-denoting nouns

Study 1: Embeddings and categorization

Jennifer Sikos and Sebastian Padó

Frame Identification as Categorization: Exemplars vs Protoypes in Embeddingland.

Proceedings of IWCS, pages 295-306.
Gothenburg, Sweden, 2019.

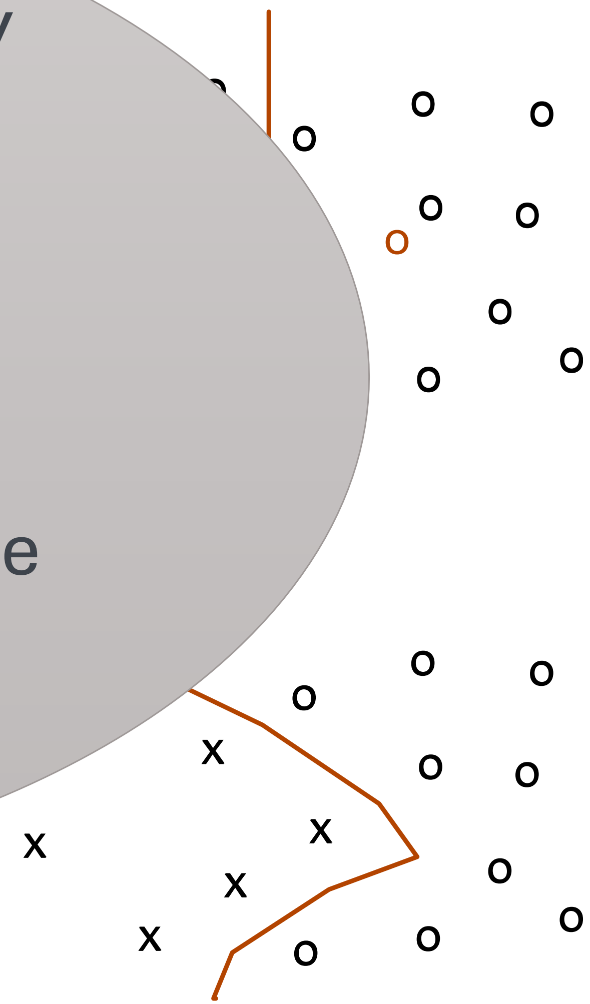
Theories of C… …tion: Dimension 1

- Prot…
- …
  - Re…
    distributi…
  - Learning … …ance storage

Prototype models support only linear decision boundaries

Exemplar models are more powerful

Prediction: Exemplar models are better if we have enough data points



Examples in Lexi

- Models of
  - *catch*
- Proto
  Erk &
- Exempla
  Erk & Padó 2
- Result: Exemplar m el better if parameters chosen well

It's not either-or:

Multi-prototype models
(Reisinger & Mooney 2013)

Theories of C… …nsion 2

- Bottom…
- **Bott…**
  by …
  - Ur…
  - Em…
- **Top-do…**
  by needs of t…
  - Supervised learn…
  - Embeddings: fine-…ng

Where are embedding-based classification models located in this space?

Does it pay off to change this?

Task: Frame Identification / Assignment

- Label a predicate instance with its frame
  - Frame Semantics: Theory of meaning based on reference to **situational category (frame)** and ability to realize its **participants**
  - Resource: Berkeley FrameNet
- Task type: Lexical disambiguation in context

	Frame: AWARENESS	
Def.	A COGNIZER has a piece of CONTENT in their model of the world.	
Semantic Roles	COGNIZER	**Peter** knows the situation. **Pat** believes that things will change.
	CONTENT	Peter knows **the situation**. Pat believes **that things will change**.
LexU.	aware.v, believe.v, comprehend.v, conceive.v, imagine.v, know.v, belief.n, consciousness.v, hunch.n, suspicion.v, conscious.a, knowledgeable.a, ...	

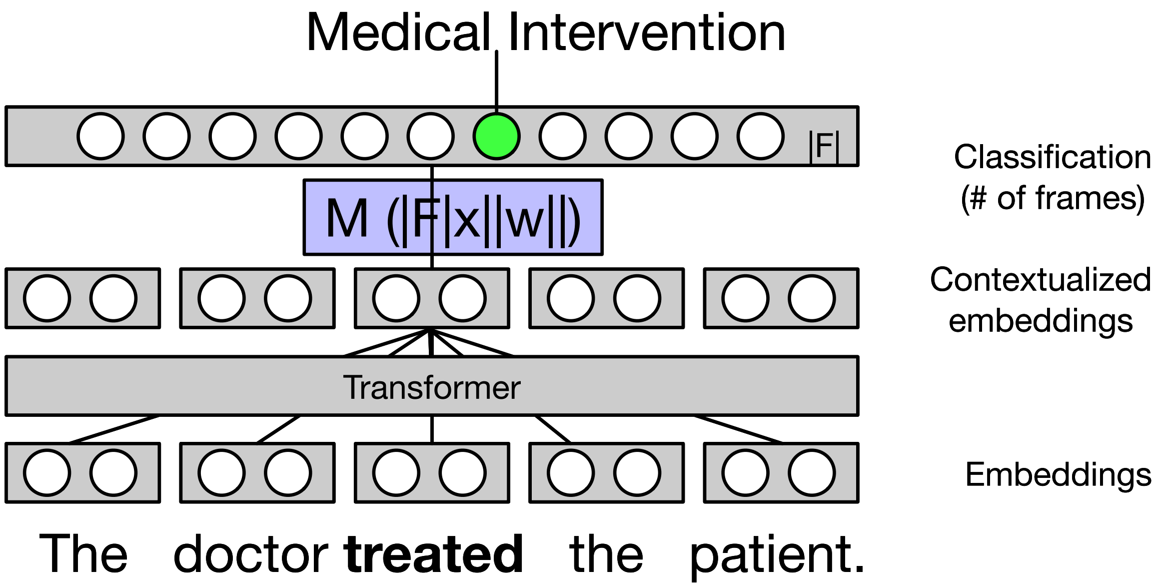
Research Hypotheses

- Let's formulate models for frame identification as
  { exemplar | prototype) x {bottom-up | bottom-up+top-down}

- **Hypotheses:** We expect…
  - …that exemplar models perform better
    (if we have enough data)
  - …that top-down + bottom-up models perform better
    (But not clear by how much)

Standard neural model

- Start: word embeddings
- Transformer contextualizes word embeddings
- Finally, predict frame

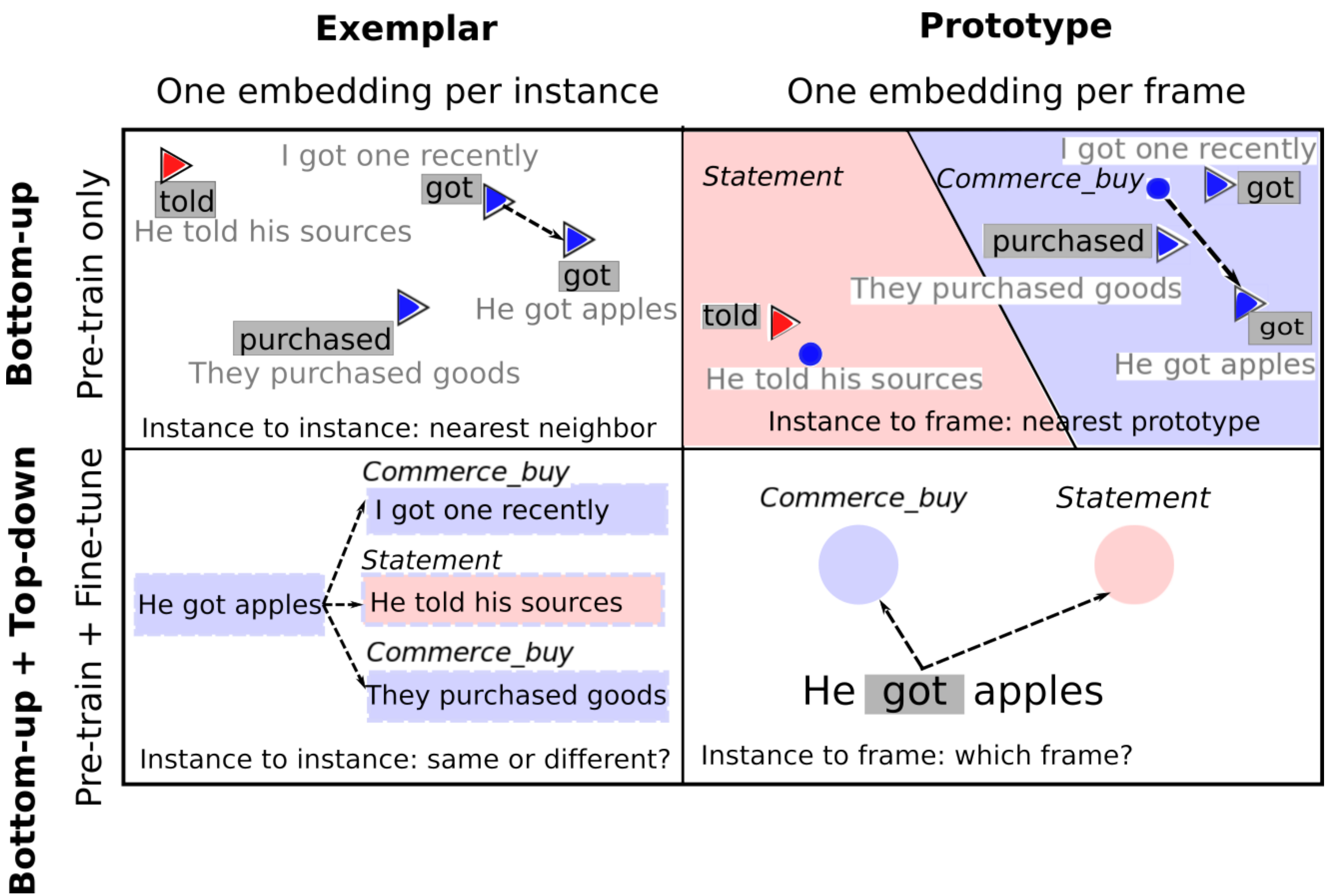


- Classification via Softmax:

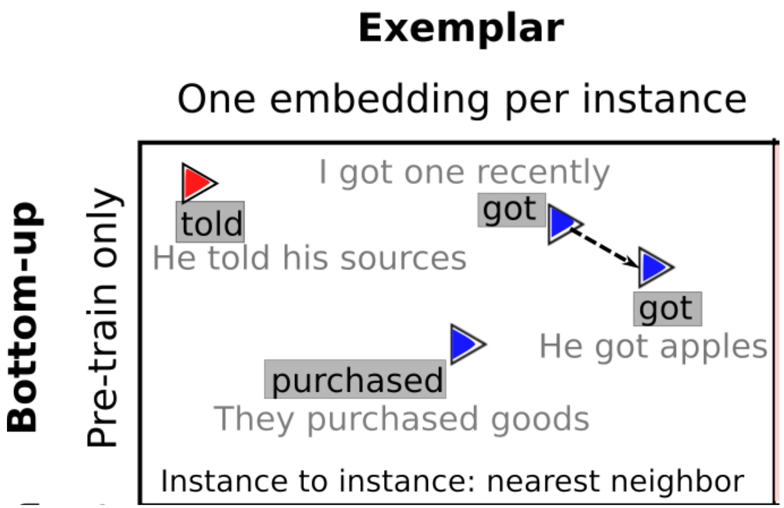
$$P(y = f|\vec{\tilde{w}}) \propto e^{\vec{\tilde{w}}^\top \cdot \vec{x}_f}$$

  - This is a prototype model
  - Frame f represented as embedding x_f in final weight matrix M

Experimental setup



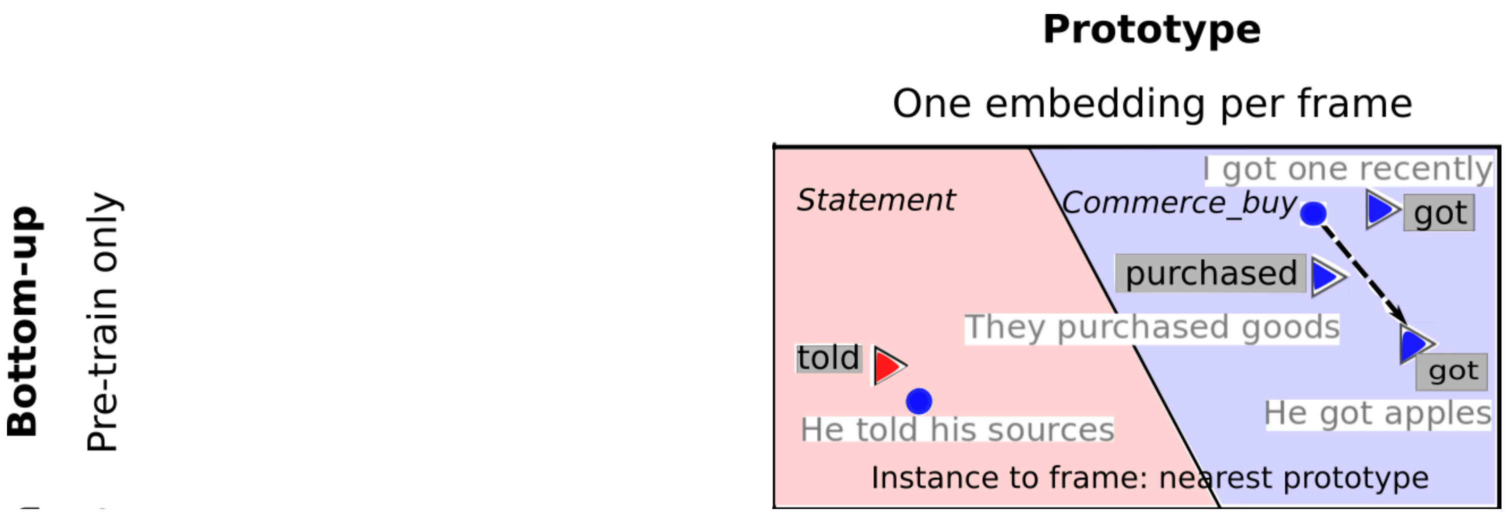
Bottom-up exemplar



- Compute embeddings for predicates in context
  - Do not use frame information (Bottom-up)

- Prediction for instance: Label of nearest neighbor instance (Exemplar)

Bottom-up prototype



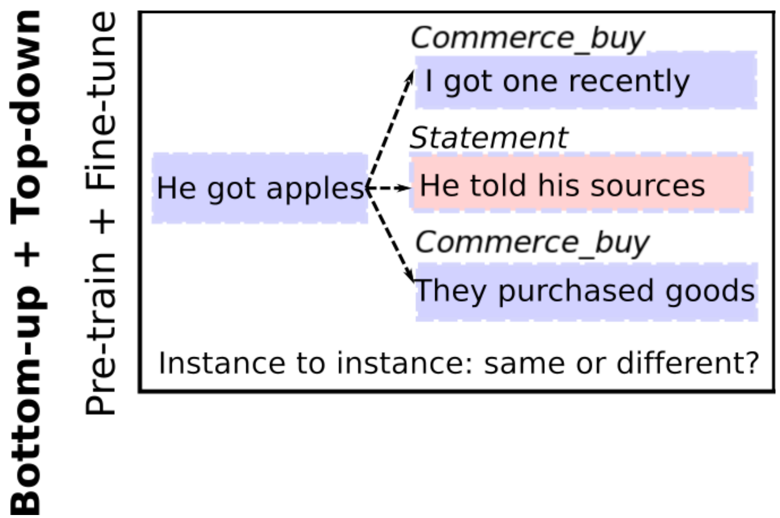
- Compute embeddings for predicates in context
  - Do not use frame information (bottom-up)
  - Aggregate into frame prototypes

- Prediction for instance: Label of nearest prototype

BU+TD exemplar

- Compute embeddings for predicates in context
  - Use frame information to fine-tune embeddings
  - Same frame/different frame classifier

- Prediction: Label of nearest neighbor instance

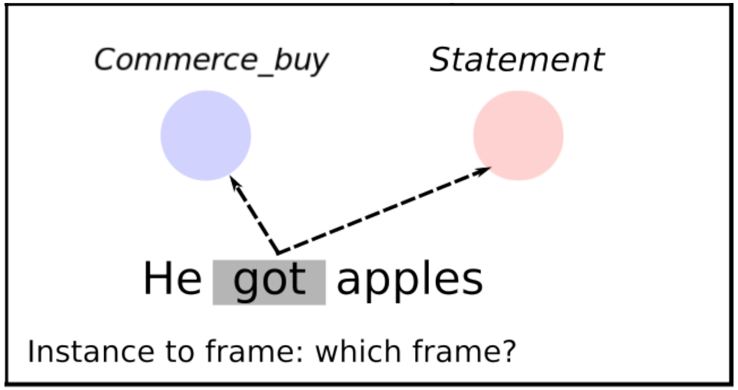


BU+TD prototype

- We have already seen this: standard case
  - Optimize classifier to predict frames for instances

Bottom-up + Top-down

Pre-train + Fine-tune



Data

- FrameNet Release 1.5 full-text annotations (Das et al. 2014)
  - ~1.2k frames
  - 78 documents (BNC)
  - Training/development: ~20k predicates
  - Testing: ~4k predicates
- We treat predicates as known
- Evaluation measure: Accuracy
  (% of predicates labelled with correct frame)

Results at global level

	Model	Full Lexicon	Ambiguous	Rare	Unseen
Our Work	Bottom-up Exemplar	82.52	64.44	81.09	11.07
	Bottom-up Prototype	84.67	69.18	83.68	09.59
	Bottom-up + Top-down Exemplar	84.09	65.06	84.18	18.89
	Bottom-up + Top-down Prototype	**91.26**	**80.77**	**91.85**	30.20

- Results were state-of-the-art (at the time)
  - Mostly due to use of recent embedding model BERT(-large)
- Bottom-up+top-down works best – unsurprisingly
- But: only significant improvement for prototype models
  - Prototype model profits much more from top-down tuning

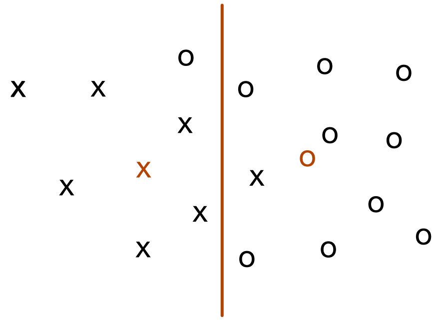
Results at frame level

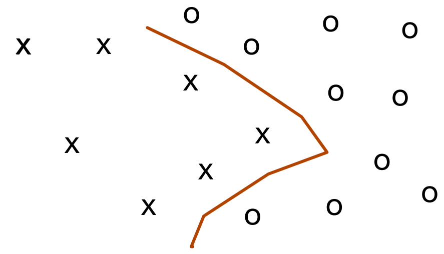
Frame	**BU+TD Prototype**	BU+TD Exemplar	BU Prototype	BU Exemplar
CAPABILITY	1.00	0.73	0.48	0.73
POSSESSION	1.00	0.94	0.92	0.81
WEAPON	1.00	0.97	0.98	1.00
LOCATIVE_RELATION	0.97	0.84	0.89	0.79
TEMPORAL_COLLOCATION	0.89	0.76	0.76	0.71

- Some frames very hard to get right
- Common denominators: *ambiguity* and *abstractness*
  - Verb *can* evokes PRESERVING, CAPABILITY, LIKELIHOOD, POSSIBILITY
  - Distinctions not well represented in non-fine-tuned embedding space

Our Interpretation of Study 1

- Traditional benefit of exemplar models: non-linear decision boundaries





  - Implicit assumption: Representations are (largely) fixed
- Neural networks models do representation learning
  - If boundary is not linear, fine-tuning will make it linear
  - Then, a prototype model is sufficient
- Exemplar model could do same, but is hampered by sparsity in this experiment (~5 instances/class): your mileage may vary!

Study 2: How to build word embeddings

Matthijs Westera, Abhijeet Gupta, Gemma Boleda and Sebastian Padó

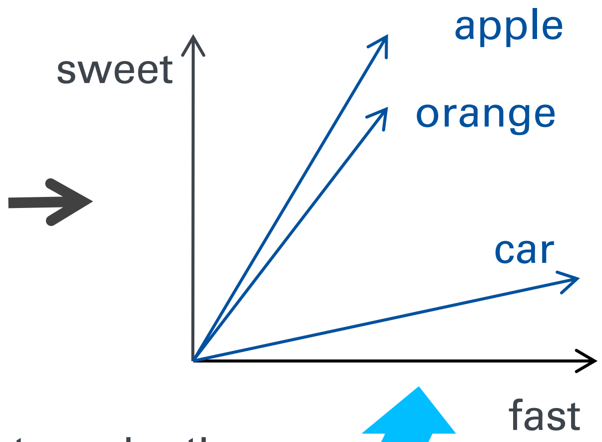
Distributional models of category concepts based on names of category members.

Cognitive Science, 45(9):e13029, 2021.

Category representation

- Fairly common assumption in distributional semantics: Embedding of **word x** is representation of **category X**
  - Our focus: nouns – “noun-based model”

	sweet	fast
orange	5	1
apple	4	1
car	1	20



Categorization:
instances → senses

Categorization:
senses → semantic classes

Is this reasonable?

- Implicit assumption: Instances of a word contribute informatively towards denoted category. True?

- Word sense — **Elephant** in the room

- Informativity — **Grass** is green

- Speaker intent — **Fotograf** vs. **Fotografin** (generic/male vs. female photographer)

Could also be framed as exemplar model.
We did not investigate this -- sparsity.

- Categor[…] „public“ entities that instanti[…]em
  - Category = prototype of embeddings of named entities

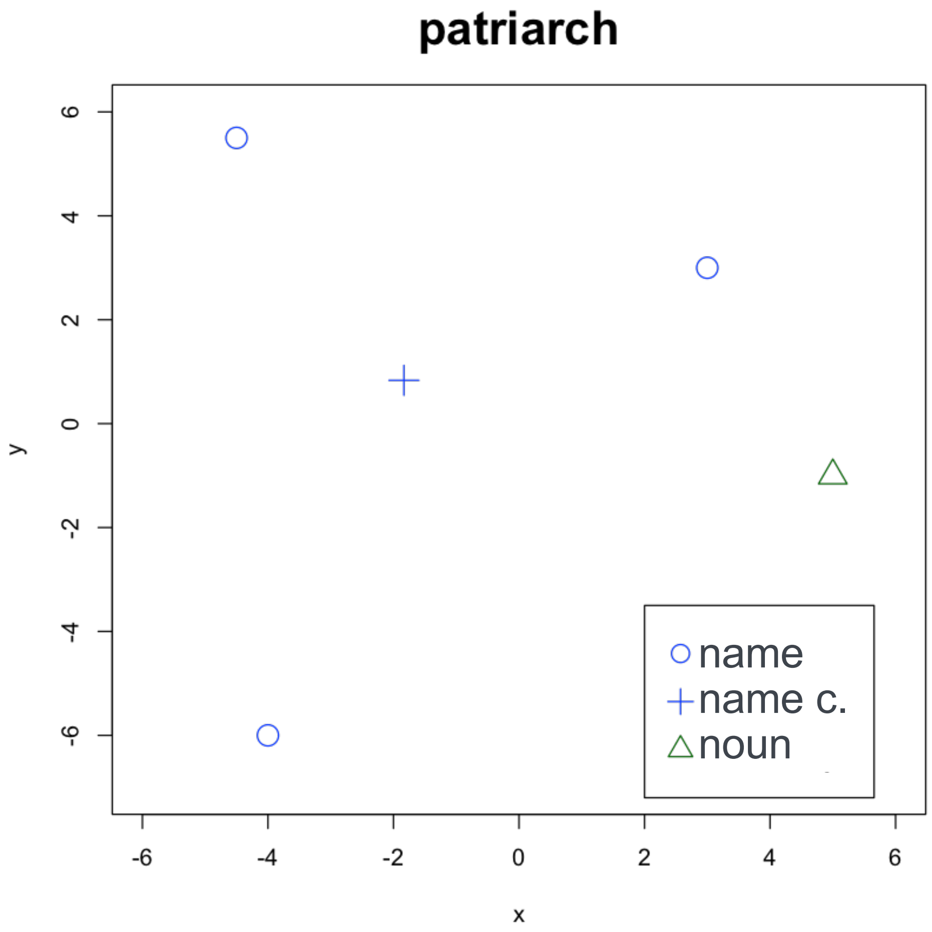
$$\overrightarrow{\textsc{lawyer}} = \overrightarrow{\text{W. Jennings}} + \overrightarrow{\text{A. G. Hays}} + \ldots$$

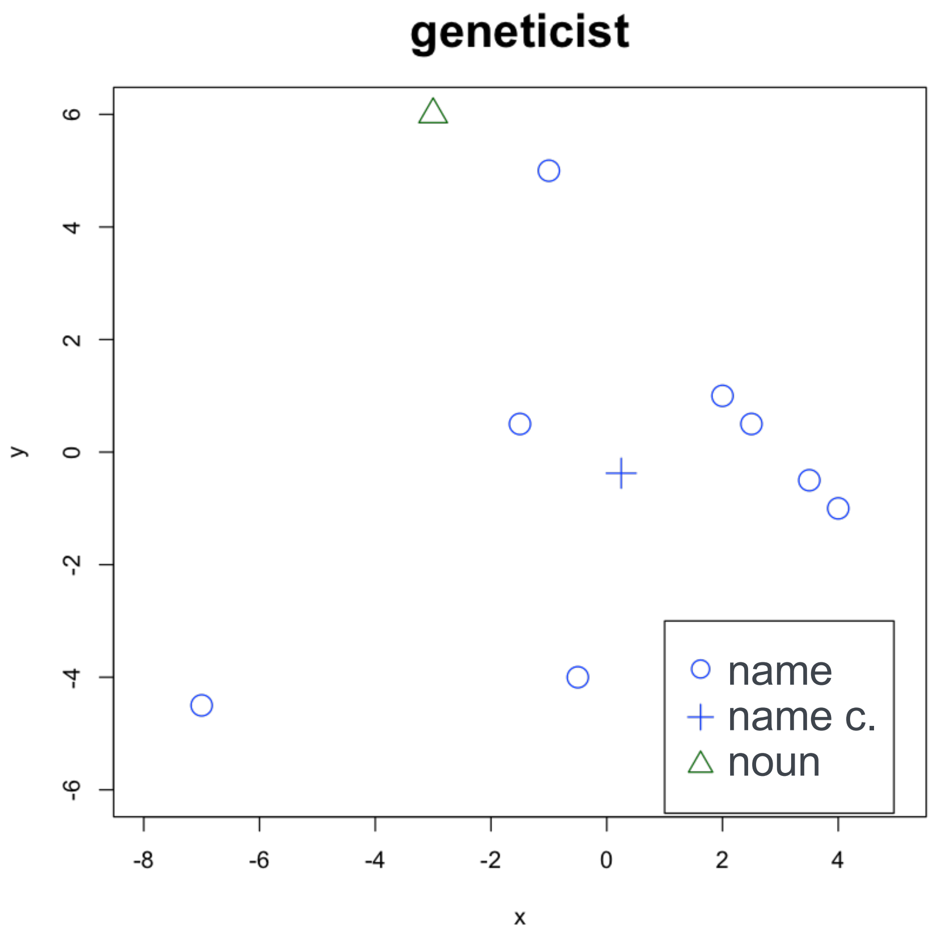
$$\overrightarrow{\textsc{artist}} = \overrightarrow{\text{S. Dali}} + \overrightarrow{\text{J. S. Bach}} + \ldots$$

$$\text{rel}(\textsc{lawyer}, \textsc{artist}) = \cos(\overrightarrow{\textsc{lawyer}}, \overrightarrow{\textsc{artist}})$$

  - Conceptual advantages over the noun-based model:
    - Instances more properly „category members“ than mentions of the category noun
    - Names are rigid designators: less interference from pragmatics

Names, nouns are used differently





Research Question and Hypothesis

- Question: How well do name-based embeddings do in accounting for category-related knowledge?
  - Experiment 1: Category relatedness
  - Experiment 2: Category membership

- Hypothesis: Better than noun-based embeddings

A Data for Instantiation

- Set of pairs (instance, category)
- Source: WordNet
  - Pairs: instance hypernym relation
  - Categories are synsets

Domain	No. of Pairs	No. of Entities	No. of Cats	Example pair
Person	2408	2076	98	Emmy Noether, mathematician
Location	1665	1436	26	Oaxaca, city
Object	547	546	18	Nile, river
Communication	48	48	5	Hail Mary, prayer
Artifact	45	45	3	Cornell, university
Act	43	43	4	Alamo, siege
Other	34	34	5	Paleocene, epoch
Total unique	4790	4180	159	

Experiment 1: Category relatedness

- Research question: Which model can account better for human judgments of category relatedness?
  - Turns out there is no dataset for **category** relatedness
- Build one via crowdsourcing experiment
  - Ranking task for category pairs
  - About 50/50 within-domain and across-domain pairs

Which pair of categories are more related to each other?

1. wheel ↔ car

2. building ↔ crane *(type of bird)*

Experiment 1: Embedding computation

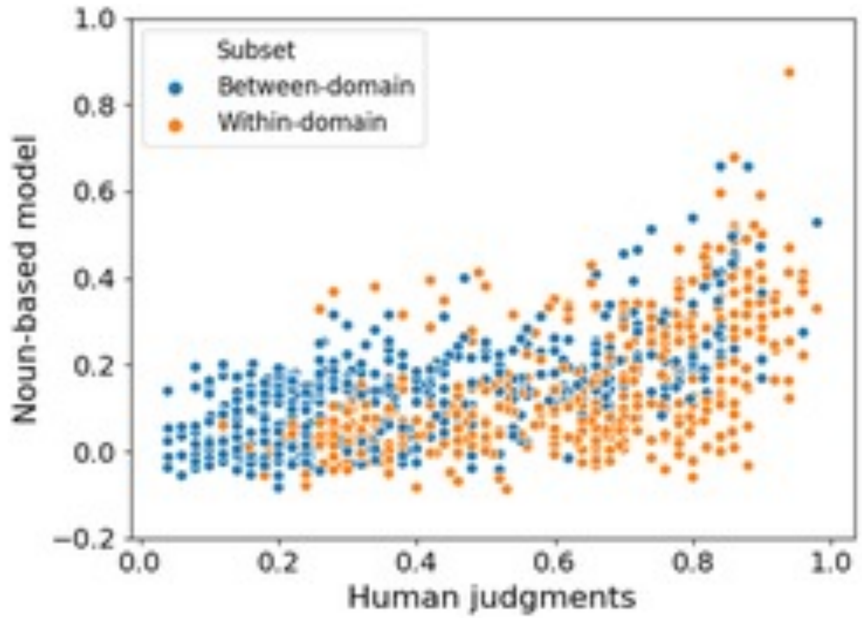
- Embedding space: static Google News embeddings

- Noun-based model
  - Category = noun embedding
- Name-based model
  - Obtain list of **n** names from our resource
  - Category = prototype of name embeddings

Experiment 1: Model comparison

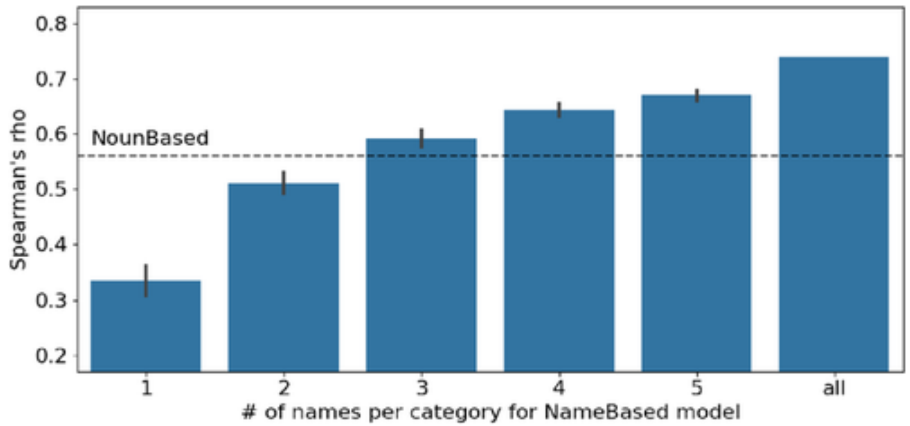
- Task: Predict relatedness for category pairs
  - Evaluation: Ranking correlation (Spearman)

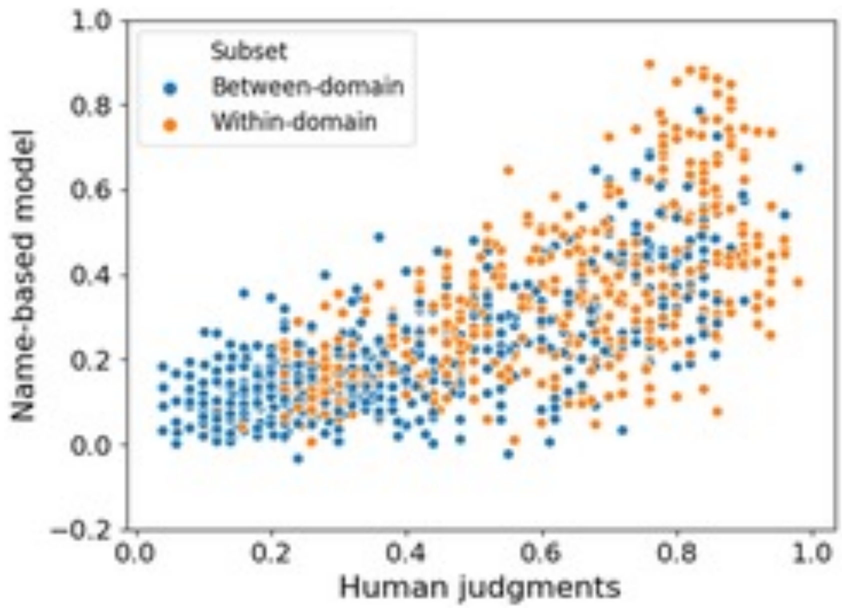
- Noun-based model:
  Performance: $\boldsymbol{\rho} = 0.56$



Experiment 1: Model comparison

- Name-based model
  - Performance improves with number of names
  - Better than noun-based model from 3 names





- Best model at $\rho = 0.74$
  - Improvement in particular on within-domain judgments

Experiment 2: Category membership

- Task: Given a pair of embeddings, is E1 instance of E2?
- Balanced evaluation dataset:

entity-match cat.	Göteborg-town
entity-mismatch cat.	Göteborg-cat
cat.-entity	town-Göteborg
entity-entity	Göteborg-Oslo

- Mismatching categories drawn within- or across-domain
- Decision architectures
  - Weakly supervised option : Cosine between embeddings
  - Fully supervised option: 1 hidden layer + classification layer
- Same embeddings as in Exp 1: Noun-based, name-based

Experiment 2: Results (more in paper)

		Noun-based		Name-based	
Dataset	**BL**	**Cos**	**NN**	**Cos**	**NN**
Within- and across-domain confounders	0.25	0.43	0.74	0.59	**0.85**
Within-domain confounders	0.25	0.41	0.51	0.55	**0.76**

- Cos not great: representation important
- Name-based model substantially better than noun-based
- Particular improvement for within-domain confounders

Our Interpretation of Study 2

- Category nouns are suboptimal cate
  - Ambiguity, pragmatics influencing us
- Prototypes of instantiating entities w
  - But why?

    Syntactic effect (CN vs. PN)?

    Semantic type effect (entity/category)?

    Specificity effect?
- NB. Our model requires „public named entities“
  - Not many per concept, but don‘t exist for many concepts

Comparison experiments e.g. with super/subordinate categories?

Last Words

- Much momentum in NLP has recently come from ML
  - Good ML goes hand in hand with domain understanding
- The domain of language fundamentally incorporates categories and categorization
  - It is worth examining our models from this angle
- That being said, in ML simplicity often wins (→ scaling)
  - Encoders/LLMs: next-word prediction/masked token objective
- If we want better models, we need to ask if/how these ideas can be integrated in (on top?) current approaches

University of Stuttgart
Institut für Maschinelle
Sprachverarbeitung

Thank you!

Sebastian Padó

e-mail pado@ims.uni-stuttgart.de

phone +49 (0) 711 685-81400

www. ims.uni-stuttgart.de/~pado

University of Stuttgart

Data & implementations should be available for
all studies – see papers or ask me