Less logical than we think: a distributional interpretation of quantifiers

Aurélie Herbelot

University of Trento
Centre for Mind/Brain Sciences

Geneva 2016

Introduction

The logical meaning of quantifiers

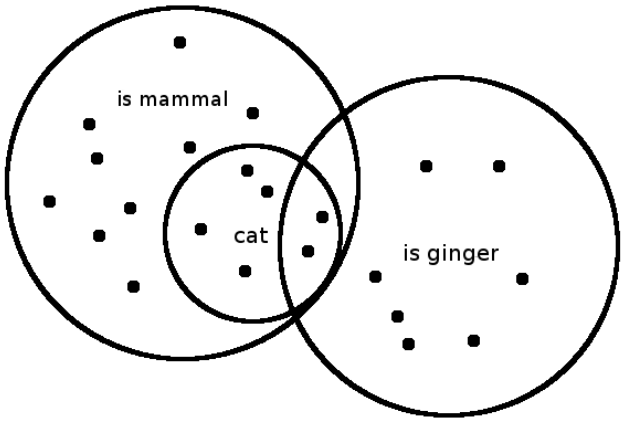
- Aristotle's 'square of opposition'

	Affirmation	Denial
Universal	Every A is B	No A is B
Particular	Some A is B	Not every A is B

- Modern formal logic:
  - $\exists$: $\exists x[cat'(x) \wedge sleep'(x)]$
  - $\forall$: $\forall x[cat'(x) \longrightarrow sleep(x)]$

Generalised quantifiers

- Quantifiers have a restrictor and a scope.
  All cats are mammals. Some cats are ginger.
- Simple interpretation: set overlap.



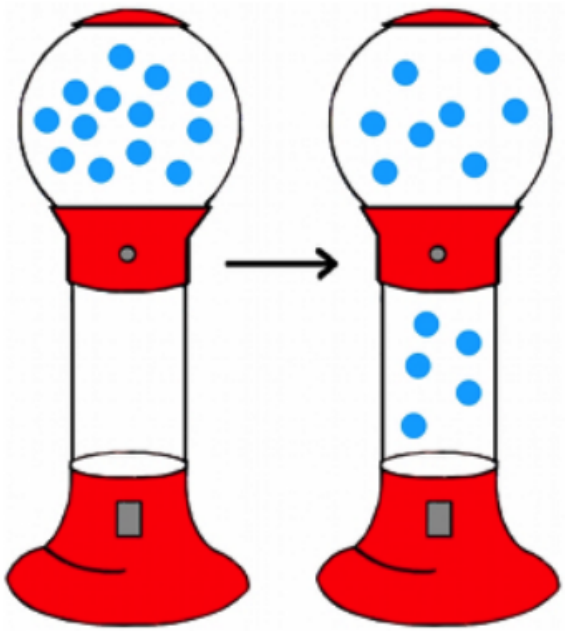
- The logic selects *individuals* over which to quantify:
  $\exists x$, $\forall x$, etc.

Beyond $\exists$ and $\forall$

- *no*: monotone decreasing.
- *most*: what is *most*? More than half? Nearly all?
- *many*: *Many cars have a GPS*, *Many dogs have three legs*.
- *the, a*: *The cat sleeps*, *The cat is a mammal*, *A cat sleeps*, *A cat is independent*, *Have you fed the fish?*
- $\emptyset$: generics (Carlson 1977). *Cats are mammals*, *Ducks lay eggs*, *Mosquitoes carry malaria*.
- ...

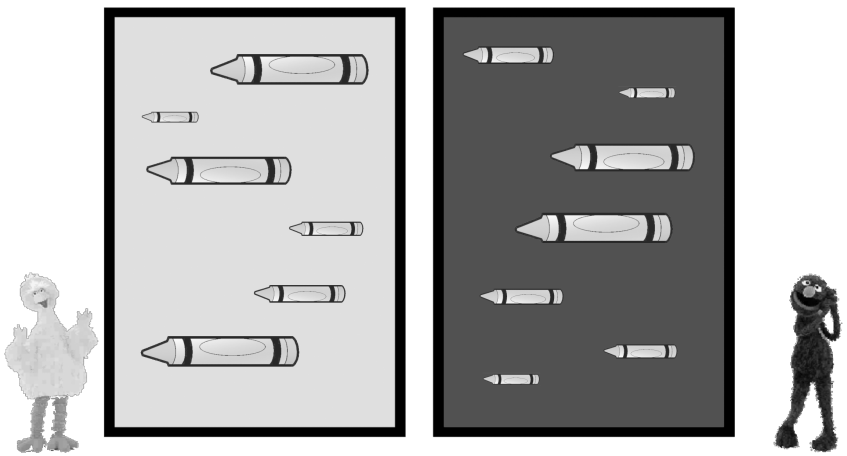
The pragmatics of quantifiers

- Some quantifiers 'feel better' than others.
- Gumball machine experiments: *You got two/some/some of/all (the) gumballs* (Degen & Tanenhaus 2015).



The psychology of quantifiers

- Children acquire quantifiers *after* generics (Hollander et al 2002).
- Children acquire numerical abilities (counting) *after* the Approximate Number Sense (ANS) (Mazzocco et al 2011).



"Who has more crayons?"

- Adults make quantification 'mistakes':
  (All) ducks lay eggs. (Leslie et al 2011).

Non-grounded quantification

- *All cats are mammals, Most cats have four legs, We had profiteroles for dessert (at the restaurant last night).*
- In non-grounded quantification, it is often unclear what exactly the restrictor's set consists of. E.g. no one knows the exact composition of the set of cats.
- Often, the set will anyway be too large to count: *Most ants have six legs.*

The obvious question

Do we need the x in $\exists x$?

The obvious question

Perhaps not...

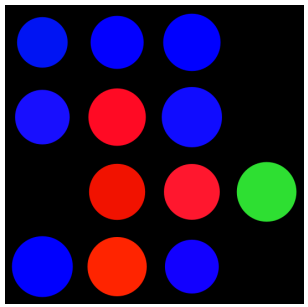
This talk

- A different take on quantifiers...
- What is a model without (sets of) individuals?
- Where does such a model come from?

Quantifying over visually grounded information (Ongoing work with Ionut Sorodoc et al)

Counting or not counting?

- The relation between quantification and counting is unclear.
- Example: *All circles are blue. True or false?*
  - *False: there are circles that are not blue.* (The cardinality of the restrictor doesn't matter.)
  - *False: there are 12 circles, 5 are not blue.* Standard set-theoretic interpretation, which allows for similar treatment of *Five out of twelve circles are not blue.*



Counting or not counting?

- Train a neural network to perform a simple quantification task: *Are no/some/all circles blue?* (The restrictor is fixed, the network learns the quantification of properties.)
- **Setup 1:** the network is given explicit cardinality information. E.g. there are 7 blue circles, 4 red circles and 1 green circle. Three deterministic rules must be learnt.
- **Setup 2:** the network is explicitly *prevented* from counting. Rather, it is made to build an aggregate vector of the image.
- 5000 randomly generated images, split into training, validation and test set (70%, 10% and 20%).

Setup 1

- The input to the network is the concatenation of a property frequency vector and a one-hot vector defining the property to quantify over.
- 7 blue circles, 4 red circles and 1 green circle.
  Query: *red*.
  Desired answer: *some*.
- Linear transformation + ReLU activation + softmax.

Setup 2

- A memory network (also tested RNNs).
- Each colour in the grid is represented by a vector. We only considered colours under a certain similarity threshold (< 0.70) to avoid confusion.
- Some Gaussian noise is added to each individual vector to account for potential visual variations in the properties under consideration.
- The input is the 16 vectors corresponding to the circles in the image.
- Aggregation: dot product of memory and 'query', which results in a weighted average of the memory (with or without softmax).

Results

Models	**familiar**	**unseen quantities**	**unseen colours**
`Counting`	71.8	80	33.4
`qMN`	**87.7**	**97**	**59.1**

Table: Model accuracies (in %)

Quantifying over non-grounded sets (Work with Eva Maria Vecchi)

Do people have models in their heads?

- Premise 1: people have conceptual knowledge.
- Premise 2: concepts are not sets (concept *ant* is not the set of ants).
- Is quantification derivable from concepts?
- To what extent are the resulting models shared amongst individuals?

The research question

- How do native speakers of English model relations between non-grounded sets?
- Given the generic *Bats are blind*:
  - how do humans quantify the statement? (*some, most, all* bats?)
  - what does this say about their concepts of *bat* and *blindness*?
- Problem: explicit quantification cannot directly be studied from corpora, being rare in naturally occurring text (7% of all NPs – see Herbelot & Copestake 2011).

Quantifying the McRae norms

- The McRae norms (2005): a set of feature norms elicited from 725 human participants for 541 concepts.
- The dataset contains 7257 concept-feature pairs such as:
  - *airplane used-for-passengers*
  - *bear is-brown*
- ... quantified.

Annotation setup

- Three native English speakers (one Southeast-Asian and two American speakers, all computer science students.
- For each concept-feature pair (C, f) in the norms, provide a label expressing the ratio of instances of C having the feature f.
- Allowable labels: NO, FEW, SOME, MOST, ALL.
- An additional label, KIND, for usages of the concept as a kind (e.g. *beaver symbol-of-Canada*).

Minimising quantifier pragmatics

- The quantification of *bats are blind* depends on:
  - the speaker's beliefs about the concepts *bat* and *blind* (lexical semantics, world knowledge);
  - their personal interpretation of quantifiers in context (pragmatics of quantifier use).
- We focus on what people believe about the actual state of the world (regardless of their way of expressing it), and how this relates to their conceptual and lexical knowledge.
- The meaning of the labels NO,FEW,SOME,MOST,ALL must be fixed (as much as possible!)

Annotation guidelines

- ALL: 'true universal' which either a) doesn't allow exceptions (as in the pair *cat is-mammal*) or b) may allow some conceivable but 'unheard-of' exceptions.
- MOST: all majority cases, including those where the annotator knew of actual real-world exceptions to a near-definitional norm.
- NO/FEW mirror ALL/MOST.
- SOME is not associated with any specific instructions.
- **Additional guidelines:** in case of hesitation, choose the label corresponding to lower set overlap (i.e. prefer SOME to MOST, MOST to ALL, etc).

Disclaimer

Disclaimer

- We are *not* modelling the way that speakers naturally use the determiners *some, most, all*, etc.
- We are modelling the perceived overlap between the set denoted by a noun and the set denoted by a predicate
- Fixing the labels' interpretation does not completely suppress all unwanted effects (see 'generic' trap, Leslie 2011).

Example annotations

Concept	*Feature*	
ape	is_muscular	ALL
	is_wooly	MOST
	lives_on_coasts	SOME
	is_blind	FEW
tricycle	has_3_wheels	ALL
	used_by_children	MOST
	is_small	SOME
	used_for_transportation	FEW
	a_bike	NO

Table: Example annotations for McRae feature norms.

- Participants took 20 or less hours to complete the task, which they did at their own pace, in as many sessions as they wished.

Inter-annotator agreement

- We need an inter-annotator agreement measure that assumes separate distributions for all three coders.
- We would also like to account for the seriousness of the disagreements: a disagreement between NO and ALL should be penalised more than one between MOST and ALL.
- Weighted Kappa (κ_w, Cohen 1968) satisfies both requirements:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} o_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} e_{ij}} \quad (1)$$

The weight matrix

- Weighted kappa requires a weight matrix to be set, to quantify disagreements.
- Setup 1: we use prevalence estimates from the work of Khemlani et al (2009) (after some mapping of their classification to ours).
- Setup 2: we exhaustively search the space of possible weights and report the highest agreement – under the assumption that more accurate prevalence estimates will result in higher agreement.

Prevalence estimates (Khemlani et al 2009)

Predication type	Example	Prevalence
Principled	Dogs have tails	92%
Quasi-definitional	Triangles have three sides	92%
Majority	Cars have radios	70%
Minority characteristic	Lions have manes	64%
High-prevalence	Canadians are right-handed	60%
Striking	Pit bulls maul children	33%
Low-prevalence	Rooms are round	17%
False-as-existentials	Sharks have wings	5%

Table: Classes of generic statements with associated prevalence, as per Khemlani (2009).

Results

	κ_w^{12}	κ_w^{13}	κ_w^{23}	κ_w^{A}
full				
KH09	.37	.34	.50	.40
BEST	.44	.40	.50	**.45**
maj				
KH09	.49	.48	.60	.52
BEST	.57	.53	.67	**.59**

Table: κ_w for MCRAE$_{full}$ and MCRAE$_{maj}$. Best estimates for exhaustive search are NO (0%), FEW (5%), SOME (35%), MOST (95%), ALL (100%)

Per-feature agreement

BR Label	Example	Freq.	κ_w^{12}	κ_w^{13}	κ_w^{23}	κ_w^{A}
taxonomic	axe a_tool	713	.66	.48	.56	.57
visual-form	ball is_round	2330	.48	.44	.54	.49
function	hoe used_for_farming	1489	.36	.35	.50	.40
encyclopaedic	wasp builds_nests	1361	.39	.34	.37	.37
visual-colour	pen is_red	421	.44	.27	.30	.34
visual-motion	canoe floats	332	.28	.20	.46	.31
smell	skunk smells_bad	24	.34	.48	.12	.31
taste	pear tastes_sweet	84	.22	.29	.36	.29
tactile	toaster is _hot	242	.19	.31	.30	.27
sound	tuba is_loud	143	.11	.10	.36	.19

Table: Per-feature agreement for MCRAE$_{full}$, sorted by κ_w^A

General observations

- Substantial agreement on the majority test set: humans do have similar 'models' of the world (phew!)
- Even when features are reliably produced for a given concept, their quantification may vary significantly between annotators.
- Agreement is highly dependent on the corresponding functional or sensory type.
- No wonder children acquire generics before quantifiers...
- No wonder explicit quantification is infrequent (a cause for disagreements)...

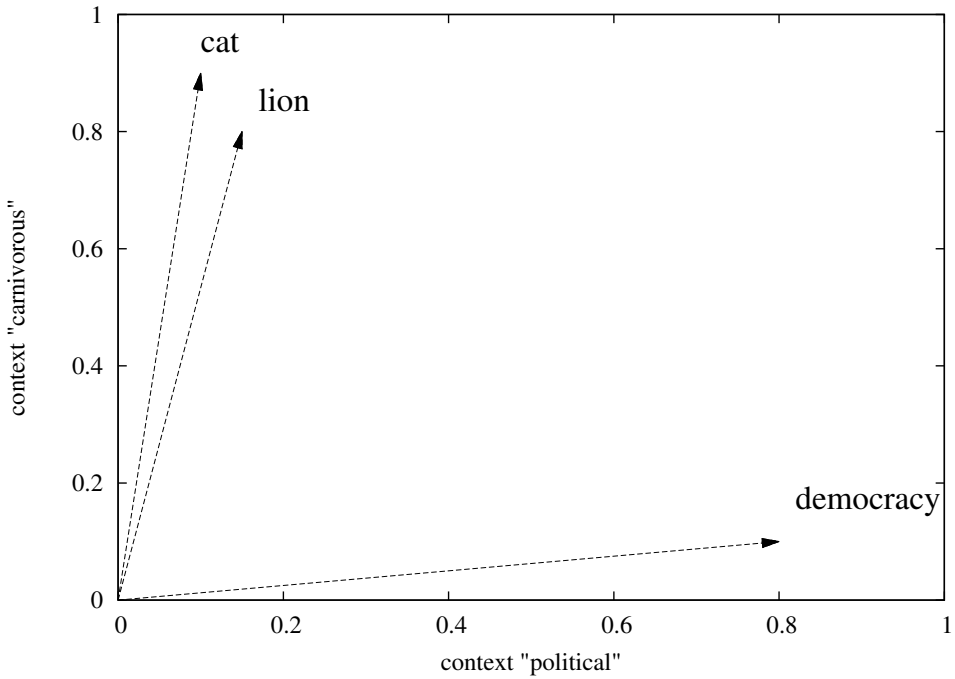
Many speakers, many worlds

- There isn't one model of the world out there. There are as many world as there are speakers. (Bad for a cognitively plausible truth-theoretic semantics.)
- Can we explain how models emerge in a speaker-dependent way?
- Can we explain how the speaker-dependent models significantly overlap?

From distributional to set-theoretic spaces (Work with Eva Maria Vecchi)

Distributional semantics

- ‘Meaning is use’.



- DS is a general representation of the usages of a word. Akin to concept representation.
- Rarely talked about: DS is by nature a theory that accommodates speaker-dependent effects.

A state-of-the-art distributional cat (Baroni et al, 2014)

0.042 seussentennial
0.041 scaredy
0.035 saber-toothed
0.034 un-neutered
0.034 meow
0.034 unneutered
0.033 fanciers
0.033 pussy
0.033 pedigreed
0.032 sabre-toothed
0.032 tabby
0.032 civet
0.032 redtail
0.032 meowing
0.032 felis
0.032 whiskers
0.032 morphosys
0.031 meows
0.031 scratcher
0.031 mouser
0.031 orinthia
0.031 scarer
0.031 repeller
0.031 miaow
0.031 sphynx
0.031 headbutts
0.031 spay
0.030 fat
0.030 yowling
0.030 flat-headed
0.030 genzyme
0.030 tail-less
0.030 shorthaired
0.030 longhaired
0.030 short-haired
0.030 siamese
0.030 english/french
0.030 strangling
0.029 sabertooth
0.029 woodpile
0.029 mewing
0.029 ragdoll
0.029 purring
0.029 whiskas
0.029 shorthair
0.029 scalded
0.029 retranslation
0.029 feral
0.028 whisker
0.028 silvestris
0.028 laziest
0.028 flap
0.028 purred
0.028 mummified
0.028 cryptozoological
...

Do cats have heads?

- grep "head" state-of-the-art-cat-distribution.txt
- 0.031179 **head**butts
  0.030823 flat-**head**ed
  0.016109 two-**head**ed
  0.009172 **head**less
- 0.002176 pilgrim
  0.002176 out
  0.002173 **head**
  0.002169 merge
  0.002165 idiot

Do cats have heads?

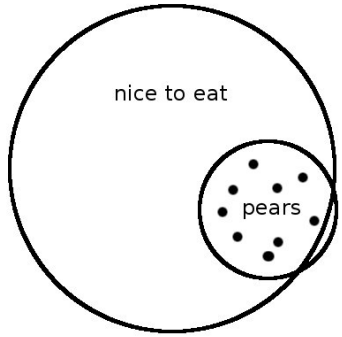
- grep "head" state-of-the-art-cat-distribution.txt
- 0.031179 **head**butts
  0.030823 flat-**head**ed
  0.016109 two-**head**ed
  0.009172 **head**less
- 0.002176 pilgrim
  0.002176 out
  0.002173 **head**
  0.002169 merge
  0.002165 idiot

Do cats have heads?

- grep "head" state-of-the-art-cat-distribution.txt
- 0.031179 **head**butts
  0.030823 flat-**head**ed
  0.016109 two-**head**ed
  0.009172 **head**less
- 0.002176 pilgrim
  0.002176 out
  0.002173 **head**
  0.002169 merge
  0.002165 idiot

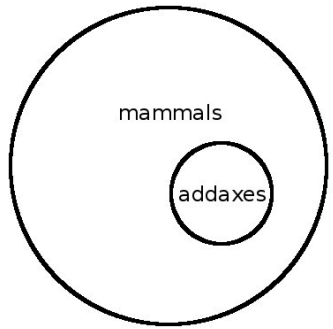
From words to worlds

I picked some pears today. They're really nice.



The reporters asked questions at the press conference.

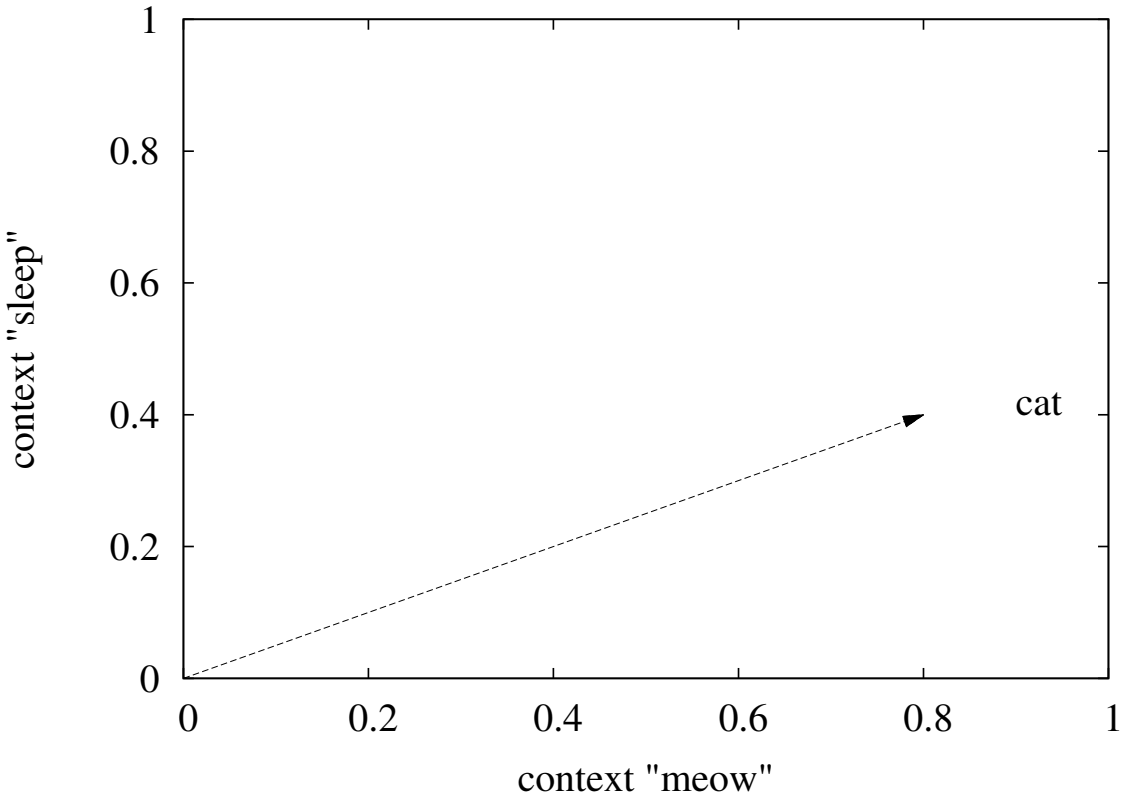
The addax is a mammal.



[Pictures: CC by beautifulcataya, NASA and Zachi Evenor.]

A set-theoretic vector space

Distributional vector spaces

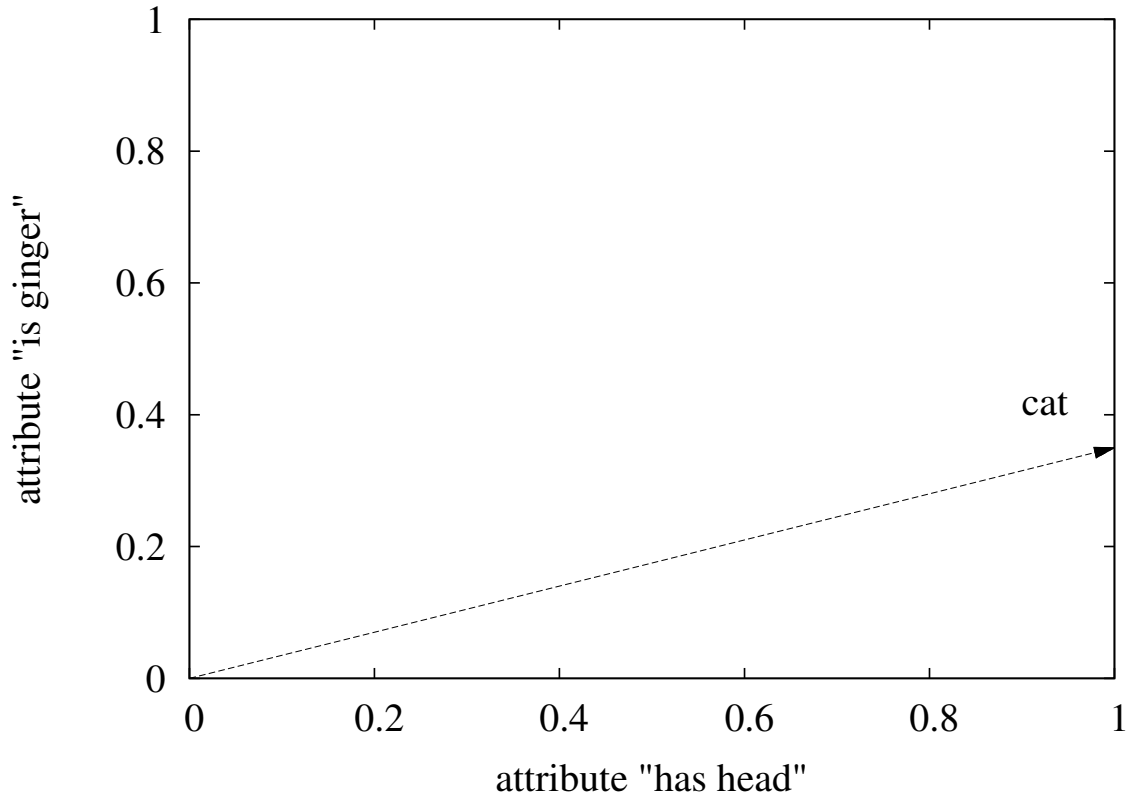


The context *meow* is very related to *cat*.

The context *sleep* is moderately related to *cat*.

Weight: how lexically characteristic a context is for a target.

Set-theoretic vector spaces



The attribute *has head* applies to ALL cats.
The attribute *is ginger* applies to SOME cats.
Weight: the set overlap between target and attribute.

QMR: The McRae norms, quantified

Concept	*Feature*	
ape	is_muscular	ALL
	is_wooly	MOST
	lives_on_coasts	SOME
	is_blind	FEW
tricycle	has_3_wheels	ALL
	used_by_children	MOST
	is_small	SOME
	used_for_transportation	FEW

Axes and hatchets

axe	*hatchet*
a tool	a tool
is sharp	is sharp
has a handle	has a handle
used for cutting	used for cutting
has a metal blade	made of metal
a weapon	an axe
has a head	is small
used for chopping	–
has a blade	–
is dangerous	–
is heavy	–
used by lumberjacks	–
used for killing	–

- Inconsistencies in McRae.
- Ideally, each concept would be annotated against all features. That is $541 * 2172 = 1,175,052$ annotations!

AD: The animal-only dataset

- Additional animal data from Herbelot (2013): a set of 72 animal concepts with quantification annotations along 54 features.
- Comprehensiveness of annotation: the 72 concepts were annotated along all 54 features. This ensures the availability of a large number of negatively quantified pairs (e.g. *cat is-fish*).

From quantifiers to weights

- Both McRae and AD datasets are annotated with natural language quantifiers rather than set cardinality ratios, so we convert the annotation into a numerical format:

$$\begin{aligned} \text{ALL} &\rightarrow 1 \\ \text{MOST} &\rightarrow 0.95 \\ \text{SOME} &\rightarrow 0.35 \\ \text{FEW} &\rightarrow 0.05 \\ \text{NO} &\rightarrow 0 \end{aligned}$$

- These weights correspond to the best weighted kappa obtained for the McRae dataset (see H&V).

Converting annotated data into vectors

Concept	*Features*	*Annotations*
hatchet	an_axe	ALL
	a_tool	ALL
	has_a_handle	ALL
	is_sharp	MOST
	is_made_of_metal	MOST
	is _used_for_cutting	MOST
	is _small	SOME

Converting annotated data into vectors

Vector	*Dimensions*	*Weights*
hatchet	an_axe	1
	a_tool	1
	has_a_handle	1
	is_sharp	0.95
	is_made_of_metal	0.95
	is _used_for_cutting	0.95
	is _small	0.35
	has_a_beak	0
	taste_good	0
	...	...

Experiments

Three configurations

Space	*# train vec.*	*# test vec.*	*# dims*	*# test inst.*
MT_{QMR}	400	141	2172	1570
MT_{AD}	60	12	54	648
MT_{QMR+AD}	410	145	2193	1595

The mapping function

- Two distributional spaces:
  - a co-occurrence based space ($\mathbf{DS}_{cooc}$ – see paper for details);
  - context-predicting vectors ($\mathbf{DS}_{Mikolov}$) available as part of the word2vec project (Mikolov et al, 2013).
- We learn a function f: $\mathbf{DS} \rightarrow \mathbf{MT}$ that transforms a distributional semantic vector for a concept to its model-theoretic equivalent.
- f: linear function. We estimate the coefficients of the function using (multivariate) partial least squares regression (PLSR).

Results

Model-Theoretic		*Distributional*		
train	*test*	DS_{cooc}	$DS_{Mikolov}$	*human*
MT_{QMR}	MT_{QMR}	0.350	0.346	0.624
MT_{AD}	MT_{AD}	**0.641**	0.634	–
MT_{QMR+AD}	MT_{QMR+AD}	0.569	0.523	–

- Results for the QMR and AD dataset taken separately, as well as their concatenation.
- Performance on the domain-specific AD is very promising, at 0.641 correlation.
- Performance increases substantially when we train and test over the two datasets (MT_{QMR+AD}).

Results

Model-Theoretic		*Distributional*		
train	*test*	DS_{cooc}	$DS_{Mikolov}$	*human*
MT_{QMR+AD}	$MT_{animals}$	**0.663**	0.612	–
MT_{QMR+AD}	$MT_{no\text{-}animals}$	0.353	0.341	–

- We investigate whether merging the datasets generally benefits all McRae concepts or just the animals.
- The result on the $MT_{animals}$ test set, which includes animals from the AD and the McRae datasets, shows that this category fares very well, at $\rho = 0.663$.
- No improvements for concepts of other classes.

Results

Model-Theoretic		*Distributional*		
train	*test*	DS_{cooc}	$DS_{Mikolov}$	*human*
MT_{QMR}	$MT_{QMR^{animals}}$	0.419	0.405	0.663
MT_{QMR+AD}	$MT_{QMR^{animals}}$	**0.666**	0.600	0.663

- We quantify the specific improvement to the McRae animal concepts by comparing the correlation obtained on the McRae animal features ($MT_{QMR^{animals}}$) after training on a) the McRae data alone and b) the merged dataset.
- Performance increases from 0.419 to 0.666 on that specific set. This is in line with the inter-annotator agreement (0.663).

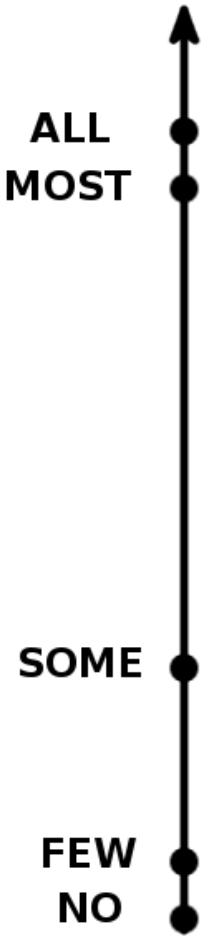
Error analysis

- Nearest neighbour analysis: the system suffers from the missing features in the QMR data.
- In the gold standard itself, some pairs are not as close to each other as they should be:

axe – hatchet	0.50
alligator – crocodile	0.47
church – cathedral	0.45
dishwasher – fridge	0.21

- Compare with *ape - monkey 0.97*.

Mapping back to quantifiers



Instance	Mapped	Gold
raven a_bird	most	all
pigeon has_hair	few	no
elephant has_eyes	most	all
crab is_blind	few	few
snail a_predator	no	no
octopus is_stout	no	few
turtle roosts	no	few
moose is_yellow	no	no
cobra hunted_by_people	some	some
snail forages	few	no
chicken is_nocturnal	few	no
moose has_a_heart	most	all
pigeon hunted_by_people	no	few
cobra bites	few	most

Producing 'true' statements with 73% accuracy.

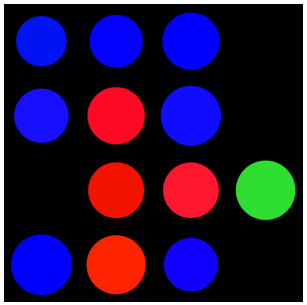
Conclusion

Contribution

0.042 seussentennial	0.032 tabby	1 walks	1 has-a_tail
0.041 scaredy	0.032 civet	1 purrs	1 has-4_legs
0.035 saber-toothed	0.032 redtail	1 meows	1 an-animal
0.034 un-neutered	0.032 meowing	1 has-eyes	1 a-mammal
0.034 meow	0.032 felis	1 has-a_heart	1 a-feline
0.034 unneutered	0.032 whiskers	1 has-a_head	0.7 is-independent
0.033 fanciers	0.032 morphosys	1 has-whiskers	0.7 eats-mice
0.033 pussy	/ 0.031 meows	1 has-paws	0.7 is-carnivorous
0.033 pedigreed	0.031 scratcher	1 has-fur	0.3 is-domestic
0.032 sabre-toothed	...	1 has-claws	...

Contribution

- Access to individuated entities is not a *necessary* condition for learning quantification.
- Similarity with the grounded problem.



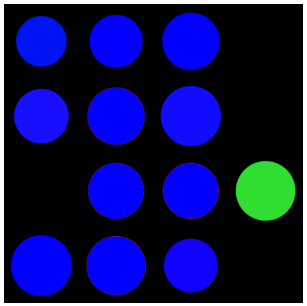
- Non-grounded quantification can be seen as an operation over concepts.
- In DS, concepts are distributionally acquired and thus speaker-dependent. Likewise, quantification can be said to be speaker-dependent (which we observed).

Have we abolished entities?

- No! Just the idea that quantification is necessarily dependent on fully specified sets.
- Entities (not sets) are hugely important:
  - Non-grounded context:
    Many computational linguists program in Python.
  - Grounded context:

Have we abolished entities?

- No! Just the idea that quantification is necessarily dependent on fully specified sets.
- Entities (not sets) are hugely important:
  - Non-grounded context:
    Many computational linguists program in Python.
  - Grounded context:



Have we abolished entities?

- No! Just the idea that quantification is necessarily dependent on fully specified sets.
- Entities (not sets) are hugely important:
  - Non-grounded context:
    Many computational linguists program in Python.
  - Grounded context:

Tomorrow...

A distributional account of entities.