Machine Learning for NLP

Reading on PLSR

Aurélie Herbelot

2018

Centre for Mind/Brain Sciences
University of Trento

Background: distributional vs truth-theoretic semantics

DS is great because...?

- Distributional Semantics (DS) allows us to build graded representations of meaning.
- Thanks to compositional distributional semantics, similarity can be calculated for any constituent, from words to sentences.
- DS models replicate not only psycholinguistic but also (to some extent!) neurolinguistic data.
- DS is so good at similarity!

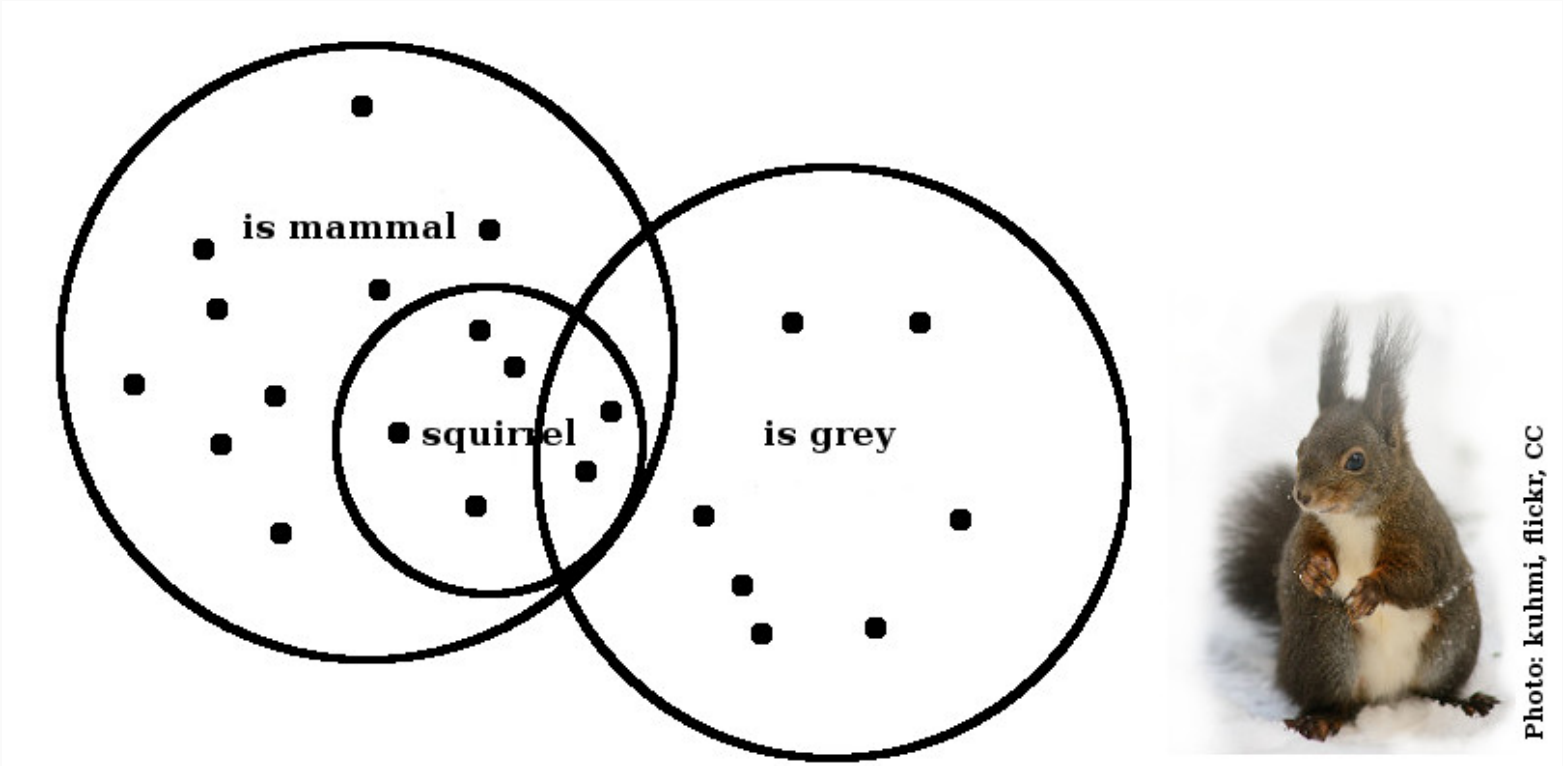
DS is great because...?

But actually...
At the theoretical level, there is nothing about DS that makes it particularly suited to modelling similarity.

Similarity is a by-product of a rich conceptual apparatus.

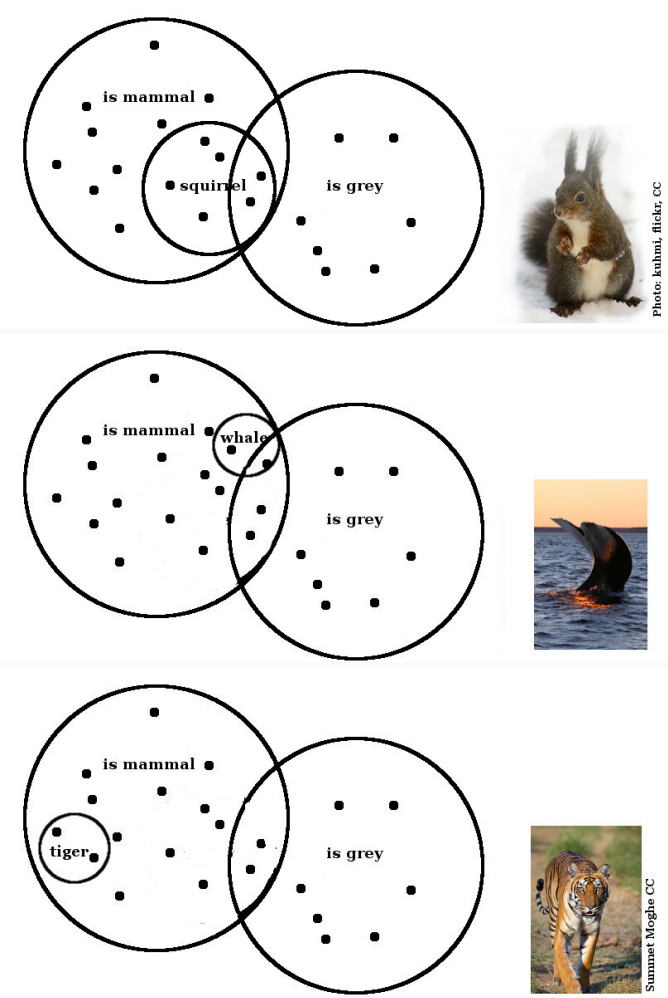
The core question is how we get our conceptual apparatus.

Model-theoretic semantics



- Truth-theoretic. It is true that *in the world*, if *x* is a squirrel, *x* is a mammal.

Model-theoretic semantics



All squirrels are mammals.
Some squirrels are grey.

All whales are mammals.
Some whales are grey.

All tigers are mammals.
No tiger is grey.

A godly model

Let's assume you are a god(dess) and have a lot of time on your hands...

You decide to write down *what there is*, starting with squirrels...

A godly model

- Squeaky is a squirrel.
- Squeaky is a mammal.
- Squeaky has claws.
- Squeaky is grey.
- Squeaky is 387 days old.
- Squeaky lives in a tree.
- Squeaky ...

A godly model

- Scott is a squirrel.
- Scott is a mammal.
- Scott has claws.
- Scott is red.
- Scott is 3 days old.
- Scott lives in a tree.
- Scott ...

A godly set of squirrels

is squirrel	256789
is mammal	256789
is grey	145675
is red	101654
has claws	256788
is 387 days old	1455
is 3 days old	1563
lives in a tree	187356
lives in the sea	0
...	

A godly set of squirrels

is squirrel	1
is mammal	1
is grey	0.57
is red	0.40
has claws	0.99
is 387 days old	0.006
is 3 days old	0.006
lives in a tree	0.73
lives in the sea	0
...	

Similarity in godly models

squirrel

is squirrel	1
is mammal	1
is grey	0.57
is red	0.40
has claws	0.99
is 387 days old	0.006
is 3 days old	0.006
lives in a tree	0.73
lives in the sea	0
...	

whale

is squirrel	0
is mammal	1
is grey	0.60
is red	0
has claws	0
is 387 days old	0.009
is 3 days old	0.016
lives in a tree	0
lives in the sea	1
...	

tiger

is squirrel	0
is mammal	1
is grey	0
is red	0
has claws	0.99
is 387 days old	0.002
is 3 days old	0.005
lives in a tree	0
lives in the sea	0
...	

So now we can do cosine (or other) similarity.

Similarity in godly models

squirrel

is squirrel	1
is mammal	1
is grey	0.57
is red	0.40
has claws	0.99
is 387 days old	0.006
is 3 days old	0.006
lives in a tree	0.73
lives in the sea	0
...	

whale

is squirrel	0
is mammal	1
is grey	0.60
is red	0
has claws	0
is 387 days old	0.009
is 3 days old	0.016
lives in a tree	0
lives in the sea	1
...	

tiger

is squirrel	0
is mammal	1
is grey	0
is red	0
has claws	0.99
is 387 days old	0.002
is 3 days old	0.005
lives in a tree	0
lives in the sea	0
...	

So now we can do cosine (or other) similarity.

Similarity in godly models

squirrel

is squirrel	1
is mammal	1
is grey	0.57
is red	0.40
has claws	0.99
is 387 days old	0.006
is 3 days old	0.006
lives in a tree	0.73
lives in the sea	0
...	

whale

is squirrel	0
is mammal	1
is grey	0.60
is red	0
has claws	0
is 387 days old	0.009
is 3 days old	0.016
lives in a tree	0
lives in the sea	1
...	

tiger

is squirrel	0
is mammal	1
is grey	0
is red	0
has claws	0.99
is 387 days old	0.002
is 3 days old	0.005
lives in a tree	0
lives in the sea	0
...	

So now we can do cosine (or other) similarity.

Similarity in godly models

squirrel	
is squirrel	1
is mammal	1
is grey	0.57
is red	0.40
has claws	0.99
is 387 days old	0.006
is 3 days old	0.006
lives in a tree	0.73
lives in the sea	0
...	

whale	
is squirrel	0
is mammal	1
is grey	0.60
is red	0
has claws	0
is 387 days old	0.009
is 3 days old	0.016
lives in a tree	0
lives in the sea	1
...	

tiger	
is squirrel	0
is mammal	1
is grey	0
is red	0
has claws	0.99
is 387 days old	0.002
is 3 days old	0.005
lives in a tree	0
lives in the sea	0
...	

So now we can do cosine (or other) similarity.

Human finitude and data sparsity

- Formal semanticists are no gods. They don't know what there is in the world. *No one* knows.
  –> Model sparsity.
- Distributional semanticists are no gods. They will never have enough data to fully describe what people *might* say about the world.
  –> Distributional sparsity.

Today: where do models come from?

- Assume humans have *some kind* of model in their heads, which allows them to utter e.g. *All cats have a heart.*
- Assume that those models are somehow acquired from the sparse data they are exposed to.
- **How can we infer models from incomplete distributional data?**

From distributional to set-theoretic spaces

A model-theoretic cat

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 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
0.029 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
...

Distributional sparsity

- 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

Distributional sparsity

- 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

Distributional sparsity

- 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

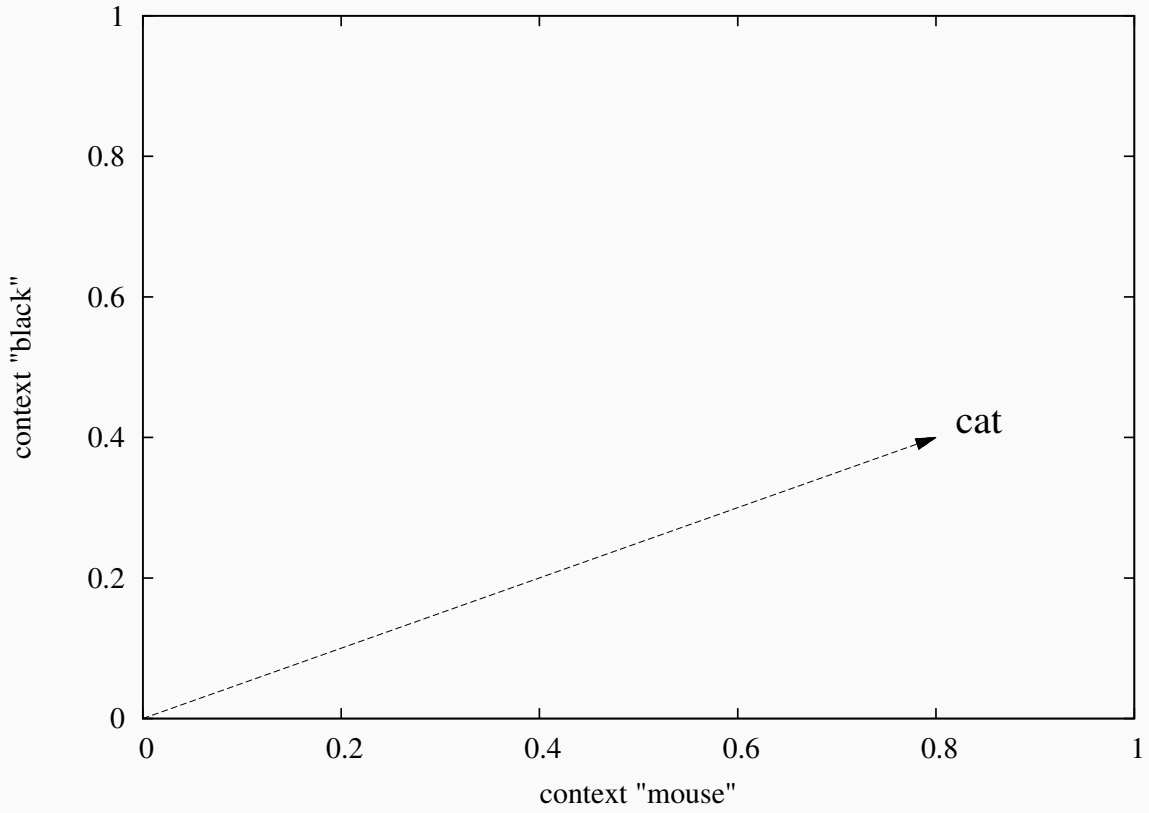
Distributional sparsity

- 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

What people say is not what there is

- Models are about things (what there is = ontological data).
- Distributional data is about things people say about things. There are a lot of things they *don't* say:
  - My cat has a head.
  - My cat, who is a mammal, is sitting on the sofa.
  - ...
- Is there a link between what people say and what there is? (Between what people say and what they believe there is?)
- Katrin Erk (2016): 'words that appear in similar contexts denote entities with similar properties'.

Distributional vector spaces

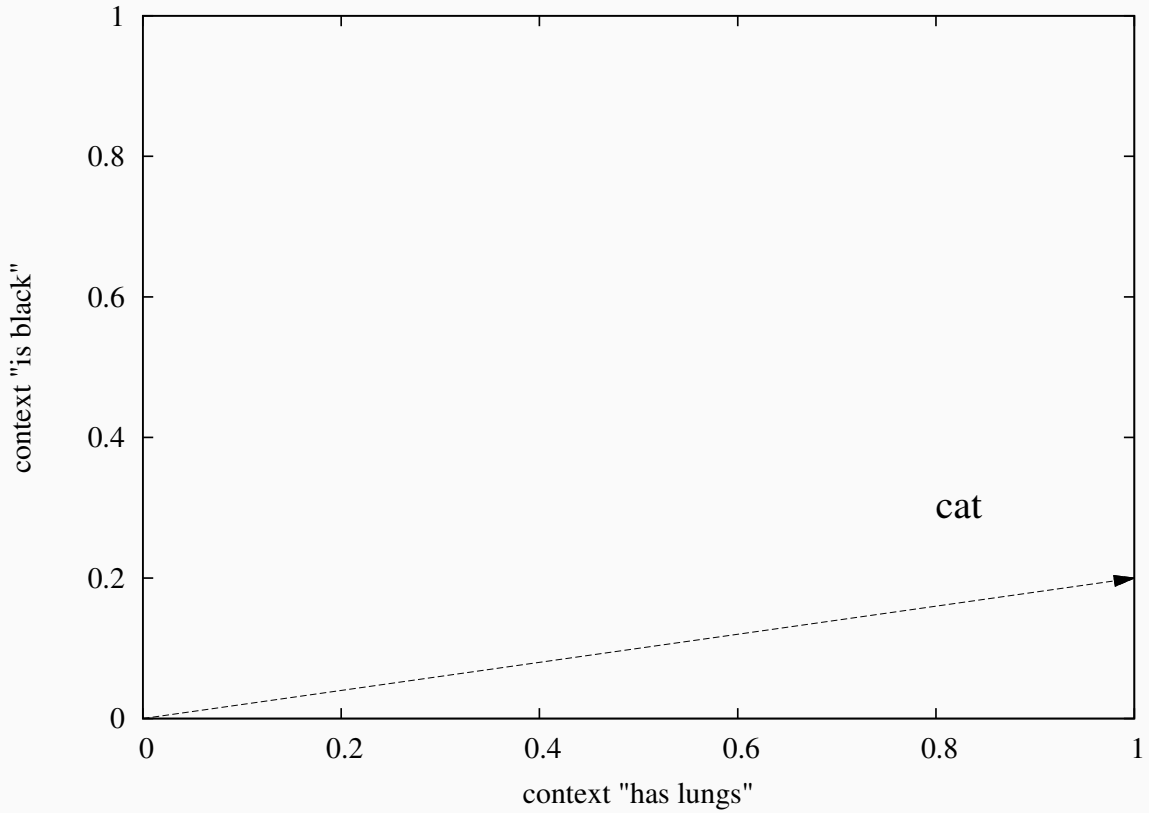


The context *meow* is very related to *cat*.

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

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

Set-theoretic vector spaces (Herbelot 2013)

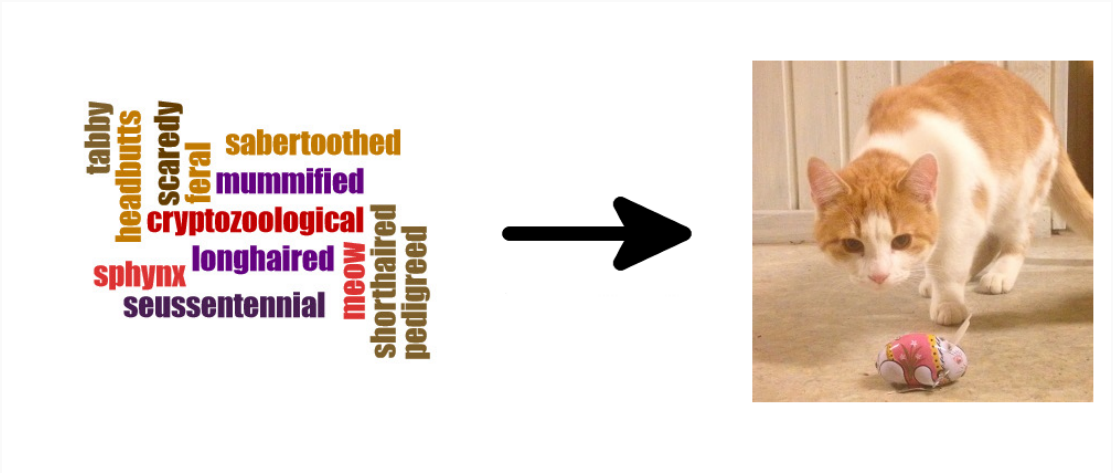


The attribute *has head* applies to ALL cats.

The attribute *is ginger* applies to SOME cats.

Weight: the set overlap between target and attribute.

Mapping from distributions to ‘models’
Herbelot & Vecchi (2015)



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
...

1 walks
1 purrs
1 meows
1 has-eyes
1 has-a_heart
1 has-a_head
1 has-whiskers
1 has-paws
1 has-fur
1 has-claws

1 has-a_tail
1 has-4_legs
1 an-animal
1 a-mammal
1 a-feline
0.7 is-independent
0.7 eats-mice
0.7 is-carnivorous
0.3 is-domestic
...

The QMR dataset (recap)

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

Table 1: Example annotations for concepts

Problem: 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 $541 * 2172 = 1175052$ annotations!

The animal-only dataset (AD)

- The McRae information is sparse.
- Additional animal data from Herbelot (2013) (henceforth AD): a set of 72 animal concepts with quantification annotations along 54 features.
- Main differences between the McRae set and AD:
  - Nature of features: the features in AD are not human elicited norms, but linguistic predicates obtained from a corpus analysis.
  - 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*).

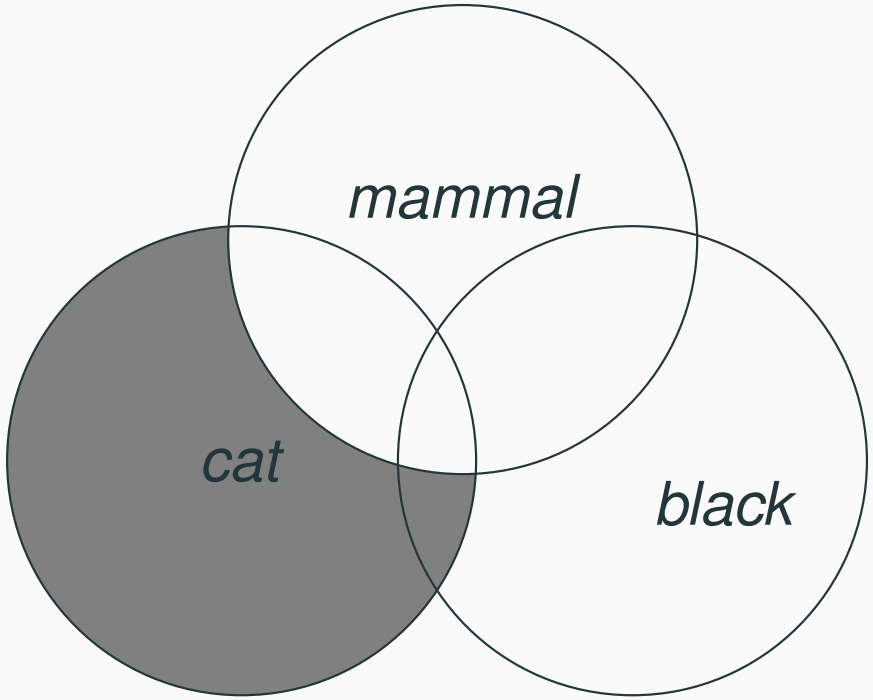
Defining a set-theoretic space

- n dimensions $d_1 \ldots d_n$ which are predicates (e.g. *is black*, *used for transportation*).
- In that space, a vector $\vec{v_c}$ (the representation of concept C) is weighted along the dimension d_k according to the ratio $\frac{|C \cap d_k|}{|C|}$.
- Weights express generalised quantifiers.

$$
\begin{array}{lll}
\textit{cat} & \{ & \textit{a\_mammal} \quad 1 \\
 & & \textit{has\_four\_legs} \quad 0.95 \\
 & & \textit{is\_black} \quad 0.2 \\
 & & \textit{lives\_underwater} \quad 0 \quad \}
\end{array}
$$

The vector CAT

- Set-theoretic equivalent of a two-dimensional vector $\overrightarrow{cat} = \{mammal\,1; black\,0.2\}$, in the form of a Venn diagram. The shaded out area signals emptiness.



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*
	an_axe	ALL
	a_tool	ALL
	has_a_handle	ALL
hatchet	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
	...	...

Experimental results

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 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	–

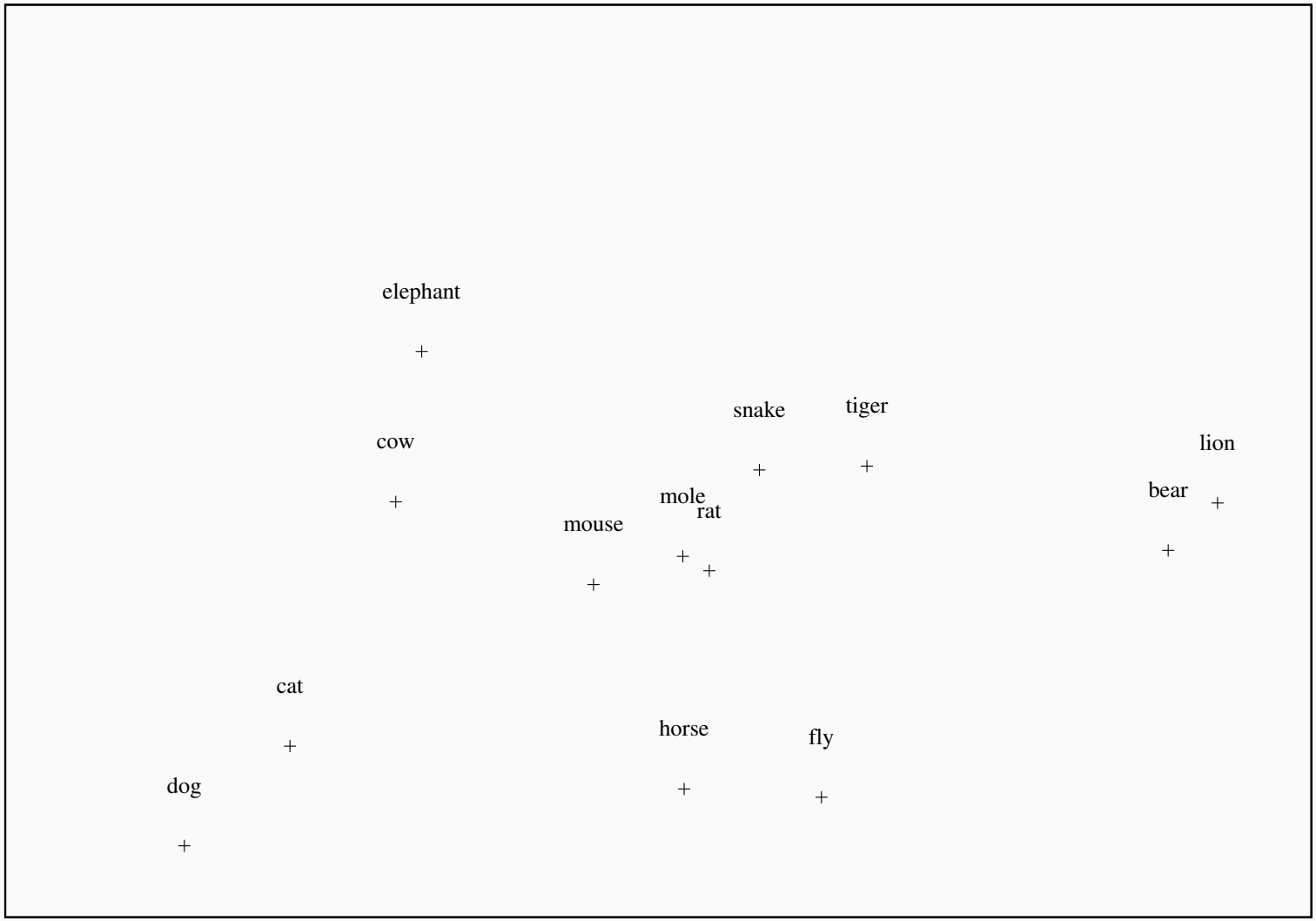
- We investigate whether merging the datasets generally benefits all McRae concepts or just the animals.
- The result on the $MT_{animals}$ test set, which includes animals from the AD and the McRae datasets, shows that this category fares very well, at $\rho = 0.663$.
- No improvements for concepts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

Nearest neighbours analysis



Nearest neighbours analysis

- For each mapped vector, get nearest neighbours in the gold standard and check whether it is close to the gold annotation.
- Example output for n=5:
  - alligator: crocodile turtle beaver otter monkey
  - chapel: building church house shed skyscraper
  - cheese: biscuit olive **cheese** parsley pear
  - marble: oak brick chandelier bookcase cupboard
  - saucer: pot pan spatula rocket skillet
  - spear: sword machete **spear** dagger revolver

Nearest neighbours analysis

	% of gold in...
top 5 neighbours	19% (29/150)
top 10 neighbours	31% (46/150)
top 20 neighbours	45% (68/150)

- Lower performance than expected, despite generally relevant neighbours.
- In many cases, the mapped vector is close to a similar concept in the gold standard, but not to itself:
  - $\overrightarrow{alligator}_{mapped}$ close to $\overrightarrow{crocodile}_{gold}$
  - $\overrightarrow{church}_{mapped}$ close to $\overrightarrow{cathedral}_{gold}$
  - $\overrightarrow{axe}_{mapped}$ close to $\overrightarrow{hatchet}_{gold}$
  - $\overrightarrow{dishwasher}_{mapped}$ close to $\overrightarrow{fridge}_{gold}$

Nearest neighbours analysis

- In the gold standard itself, some pairs are not as close to each other as they should be:

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

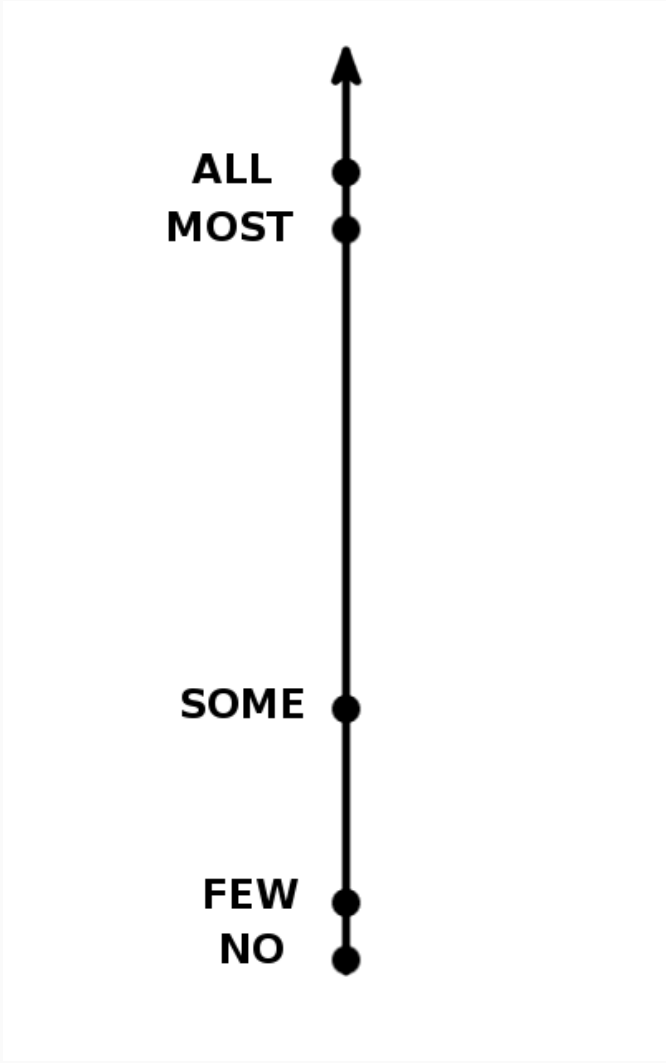
- The McRae norms do not make for a consistent semantic space because a feature that – from an extensional point of view – seems relevant to two concepts might only have been produced by the annotators for one of them.

Some set-theoretic vectors

bear	*housefly*	*plum*	*cottage*
an_animal	an_insect	a_fruit	has_a_roof
a_mammal	is_small	grows_on_trees	has_doors*
has_eyes	flies	is_round	used_for_shelter*
is_muscular	is_slender*	is_edible	a_house
has_a_head	crawls*	is_small	a_building*
has_4_legs	has_legs	has_skin	has_windows
has_a_heart	a_bug*	tastes_sweet	is_small
is_terrestrial*	stings*	is_juicy	made_by_humans*
has_hair	has_wings	has_seeds*	made_of_wood*
walks	has_eyes	tastes_good	worn_on_feet*
has_a_tail*	is_black	has_peel*	used_for_living_in
a_carnivore	is_terrestrial*	is_green*	has_rooms*
is_brown	is_large*	is_orange*	an_appliance*
a_predator	has_a_heart*	is_citrus*	has_tenants*
hunted_by_people	has_antennae*	is_yellow*	has_a_bathroom*
is_furry*	jumps*	has_vitamin_C*	has_a_kitchen*
is_large	bites*	has_a_pit	used_for_farm_equipment*
is_wooly	has_a_head*	has_leaves	found_on_farms*
has_fur	hibernates*	has_a_stem*	found_in_the_country
is_stout	is_yellow*	has_sections*	has_soles*

Example of 20 most weighted contexts in the predicted model-theoretic vectors for 4 test concepts, shown for the $\text{DS}_{cooc} \rightarrow \text{MT}_{McRae+AD}$ transformation. Features marked with an asterisk ($*$) are not among the concept's features in the gold data.

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

- It is possible to retrieve model-theoretic information from distributional data.
- But the information in the training data must be ‘complete’ enough.
- In particular, negative information (e.g. *no cats are fish*) is essential.