Compositional Distributional Semantics

Aurelie Herbelot

University of Trento
Center for Mind/Brain Sciences

Trento 2016

Outline

DS as linguistic representation

- **Composition:** find a function $f(\vec{u}, \vec{v})$ which returns the meaning of the composition of $\vec{u}$ and $\vec{v}$.
- **Lexical ambiguity:** re-weight a vector in context to get the various senses of the word it represents.
- **Inference:** if Molly is a cat, Molly is an animal, *many cats* entails *some cats*.
- Many other linguistic phenomena: mass/count distinction, relative pronouns, negation, etc, etc.

What is compositionality?

- Language is productive: from a (relatively) finite number of simple expressions (e.g. words), we can build an infinite number of novel sentences and be understood.
- Compositionality says that we have rules to combine the meaning of simple constituents to get at the meaning of the whole.

Where is compositionality?

- At the morphological level: *kind/unkind*, *do/doable*, *hero/anti-hero*...
- At the constituent level: *a cat*, *black cat*, *football match*, *long term airport car park*...
- At the sentence level:
  A cat sleeps.
  The football match was boring
  Whether or not we will stay at the long term airport car park will depend on Kim's ability to pick us up when we fly back.
- Even at the discourse level:
  Kim thinks we should go out. The cinema program looks good.
  Kim fell off the cliff. Sandy had pushed him.

Outline

Motivation

- Formal semantics gives an elaborate and elegant account of the productive and systematic nature of language.
- The formal account of compositionality relies on:
  - *words* (the minimal parts of language, with an assigned meaning)
  - *syntax* (the theory which explains how to make complex expressions out of words)
  - *semantics* (the theory which explains how meanings are combined in the process of particular syntactic compositions).

Motivation

- But formal semantics does not actually say anything about lexical semantics (the meaning of *cat*, cat', is the set of all cats in particular world).
- Distributions a potential solution?
- If we make the approximation that distributions are 'meaning', then we need a way to account for compositionality in a distributional setting.

Why not just look at the distribution of phrases?

- The distribution of phrases – even sentences – can be obtained from corpora, but...
  - those distributions are very sparse;
  - observing them does not account for productivity in language.
- Some models assume that corpus-extracted phrasal distributions are irrelevant data.
- Some models assume that, given enough data, corpus-extracted phrasal distributions have the status of gold standard.

Some distributional compositionality models

- Pointwise models: word-based model, task-evaluated.
- Lexical function model: word-based, evaluated against phrasal distributions.
- Pregroup grammar model: CCG-based model, task-evaluated.

Mitchell and Lapata (2010)

- Word-based (5 words on either side of the lexical item under consideration).
- The composition of two vectors $\vec{u}$ and $\vec{v}$ is some function $f(\vec{u}, \vec{v})$. M & L try:
  - addition $p_i = \vec{u_i} + \vec{v_i}$
  - multiplication $p_i = \vec{u_i}.\vec{v_i}$
  - tensor product $p_{ij} = \vec{u_i}.\vec{v_j}$
  - circular convolution $p_{ij} = \sigma_j \vec{u_j}.\vec{v_{i-j}}$
  - ... etc
- Task-based evaluation: similarity ratings. Multiplication is best measure. (BUT: this does't hold across all tasks!)

Example

early_j	age_n	early_j age_n
africa::9.75873	africa::3.56225	africa::34.76303
african::6.87337	african::1.88733	african::12.97231
aftermath::3.40748	aftermath::1.37812	aftermath::4.69591
afternoon::42.2096	afternoon::1.9041	afternoon::80.3712
afterwards::7.46585	afterwards::3.86807	afterwards::28.87843
again::9.00563	again::2.78339	again::25.06618
age::15.6464	age::0	age::0
aged::5.99896	aged::24.6173	aged::147.67819
agencies::4.91747	agencies::1.57129	agencies::7.72677
agency::7.28471	agency::3.13776	agency::22.85767
agent::4.63014	agent::2.24935	agent::10.41480
agents::4.21793	agents::1.68319	agents::7.09957
ages::45.003	ages::0	ages::0
ago::18.8909	ago::19.2306	ago::363.2833
agree::5.05183	agree::3.67157	agree::18.54814
agreed::6.36066	agreed::2.61272	agreed::16.61862
agreement::7.64836	agreement::0.912126	agreement::6.976268
agricultural::11.3745	agricultural::2.66057	agricultural::30.26265

Difference in top-rated contexts for *early age*

multiplication	**phrase**
1990s	talent
1980s	interested
1970s	showed
20th	learned
1960s	piano
childhood	studying
1950s	exposed
age	ages
1940s	parents
1920s	encouraged
1930s	singing
19th	educated
late	interest
century	uncle
morning	violin
stages	baronet
settlers	eldest
warning	raised

Discussion: the meaning of f

- How do we interpret $f(\vec{u}, \vec{v})$ linguistically?
- Intersection in formal semantics has a clear interpretation:
  $\exists x[cat'(x) \wedge black'(x)]$
  There is a cat in the set of all cats which is also in the set of black things.
- But what with addition, multiplication (let alone circular convolution)??

Multiplication

- Multiplication is intersective.
- But it is commutative in a word-based model:
  $\overrightarrow{\text{The cat chases the mouse}} = \overrightarrow{\text{The mouse chases the cat}}$
- Note that in a syntax-based model, things could work out:
  $\overrightarrow{\text{cat}_{subj}\ \text{chase}_{head}\ \text{mouse}_{obj}} \neq \overrightarrow{\text{mouse}_{subj}\ \text{chase}_{head}\ \text{cat}_{obj}}$

Multiplying to zero

- Multiplication has issues retaining information when composing several words. Most dimensions become 0 or close to 0:

$$\begin{pmatrix}0.45\\0.23\\0.00\\0.14\\0.76\end{pmatrix} \text{ x } \begin{pmatrix}0.11\\0.43\\0.54\\0.00\\0.39\end{pmatrix} = \begin{pmatrix}0.05\\0.10\\0.00\\0.00\\0.30\end{pmatrix} \quad \begin{pmatrix}0.05\\0.10\\0.00\\0.00\\0.30\end{pmatrix} \text{ x } \begin{pmatrix}0.00\\0.89\\0.57\\0.23\\0.42\end{pmatrix} = \begin{pmatrix}0.00\\0.09\\0.00\\0.00\\0.13\end{pmatrix}$$

Multiplying to zero

- Multiplication has issues retaining information when composing several words. Most dimensions become 0 or close to 0:

$$\begin{pmatrix}0.45\\0.23\\0.00\\0.14\\0.76\end{pmatrix} \text{ x } \begin{pmatrix}0.11\\0.43\\0.54\\0.00\\0.39\end{pmatrix} = \begin{pmatrix}0.05\\0.10\\0.00\\0.00\\0.30\end{pmatrix} \quad \begin{pmatrix}0.05\\0.10\\0.00\\0.00\\0.30\end{pmatrix} \text{ x } \begin{pmatrix}0.00\\0.89\\0.57\\0.23\\0.42\end{pmatrix} = \begin{pmatrix}0.00\\0.09\\0.00\\0.00\\0.13\end{pmatrix}$$

Addition

- Addition is not intersective: the whole meaning of both $\vec{u}$ and $\vec{v}$ are included in the resulting phrase.
- Commutativity is a problem, as with multiplication.
- No sense disambiguation and no indication as to how an adjective, for instance, modifies a particular noun (i.e. the distributions of *red car* and *red cheek* both include high weights on the *blush* dimension).
- Too much information.
- Still, in practice, simple addition has shown good performance on a variety of tasks...

Evaluation

- Similarity task at the phrase level (*AN, VN, NN*).
- Multiplication outperforms other methods.
- Results are close to human performance (which itself is not that good...) for ANs and NNs, less so for VNs.

Baroni and Zamparelli (2010)

- Word-based model for adjective-noun composition.
- Composition is the multiplication of vectors/matrices **learned** from access to phrasal distributions.
- 'Internal' evaluation: composition is evaluated against phrasal distributions.

Assumptions

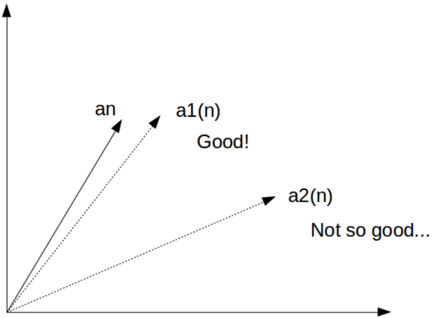
- Given enough data, distributions for phrases should be obtained in the same way as for single words.
- I.e. it is fair to assume that if we have seen enough instances of *black cat*, the context of the phrase should give us an indication of its meaning (perhaps it is more related to witches than *cat* and *ginger cat*).
- Let's say we have a vector $\vec{a}$ (*black*) and a $\vec{n}$ (*cat*), and also a $\vec{an}$ (*black cat*), we can hypothesise a composition method which combines $\vec{a}$ and $\vec{n}$ to get $\vec{an}$ (standard machine learning).

Assumptions

- There is no single composition operation for adjectives. Each adjective acts on nouns in a different way:
  - *red car*: the outside of the car is evenly painted with the colour red (visual);
  - *fast car*: the engine of the car is powerful (functional);
  - *expensive car*: the price of the car is high (abstract/relational).
- Even single adjectives will combine with various nouns in different ways:
  - *red car*: outside of the car, even paint;
  - *red watermelon*: inside of the watermelon, probably not as red as the car;
  - *red nose*: a little redder than usual, probably due to a cold.

System

- In formal semantics, adjectives are seen as functions which ‘apply’ to nouns. They take a property (a noun phrase) and return another property (another noun phrase): $A(N) = AN$.
- Test by measuring distance between a given adjective-noun combination and the corresponding phrasal distribution on unseen data.



System

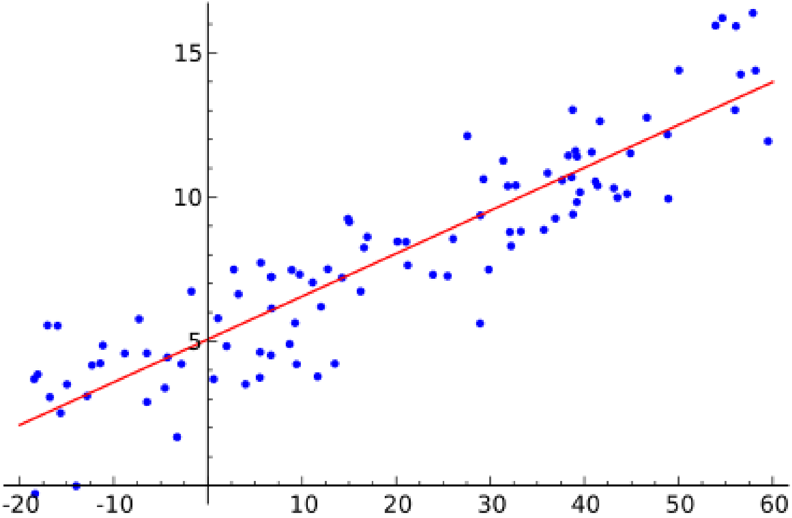
- For each adjective, a matrix is learned from actual AN phrases using partial least squares regression (PLSR).

$$\mathbf{AN} = \begin{pmatrix} a & b & c \\ p & q & r \\ u & v & w \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} = \begin{pmatrix} an_1 + bn_2 + cn_3 \\ pn_1 + qn_2 + rn_3 \\ un_1 + vn_2 + wn_3 \end{pmatrix} = \begin{pmatrix} an_1 \\ an_2 \\ an_3 \end{pmatrix}$$

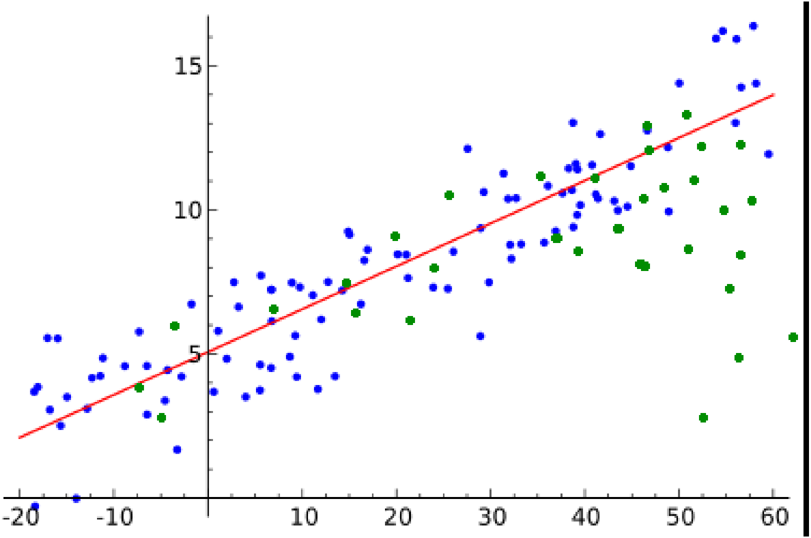
What is ‘learning’?

- Assume you have a *gold standard*: data points for which you already have the solution to the task.
- In our AN setting, the gold standard is a number of adjective-noun phrases for which we have a) a noun vector; b) an adjective vector; c) a phrase vector (e.g. $\overrightarrow{black}$, $\overrightarrow{cat}$, $\overrightarrow{blackcat}$).
- Infer a rule (in our case, a matrix) which explains the observed data.
- Check whether the rule holds for unobserved data.

Regression: training



Regression: testing



Evaluation

- Compare how close the predicted vector is to the actual, observed AN vector.
- In the original paper, the model outperforms the simple additive model.
- The lexical function has been used to model a range of linguistic phenomena. (Later today: various classes of adjectives.)

Coecke et al (2010)

- Based on pregroup grammar.
- Composition involves tensor product and point-wise multiplication.
- Evaluated on similarity task.

Thanks to Steve Clark for some of the slides!

Pregroup grammar

- A pregroup is a partially ordered monoid in which each element a has a *left adjoint* a^l and a *right adjoint* a^r such that

$$a^l \cdot a \rightarrow 1, \quad a \cdot a^r \rightarrow 1$$

- The monoid is the set of grammatical types (NP, NP^r, NP^l, NP^{rr}, NP^{ll}, S, PP, $\ldots$) with the juxtaposition operator $(\cdot)$ used to derive complex types and the empty string as unit (1)

$$NP \cdot (NP^r \cdot S \cdot NP^l) \cdot NP$$

- The composed components are vectors or matrices.

Categorial Grammar Derivation

$$\frac{Google}{NP} \quad \frac{bought}{NP\backslash S/NP} \quad \frac{Microsoft}{NP}$$

Categorial Grammar Derivation

$$
\begin{array}{ccc}
\textit{Google} & \textit{bought} & \textit{Microsoft} \\
\hline
NP & NP\backslash S/{\color{blue}NP} & {\color{blue}NP} \\
 & \multicolumn{2}{c}{\overline{}>} \\
 & \multicolumn{2}{c}{NP\backslash S}
\end{array}
$$

Categorial Grammar Derivation

$$
\begin{array}{ccc}
\mathit{Google} & \mathit{bought} & \mathit{Microsoft} \\
\hline
NP & NP\backslash S/NP & NP \\
 & \multicolumn{2}{c}{\overline{\qquad\qquad\qquad\qquad}>} \\
 & \multicolumn{2}{c}{NP\backslash S} \\
\multicolumn{2}{c}{\overline{\qquad\qquad\qquad\qquad}<} & \\
\multicolumn{2}{c}{S} &
\end{array}
$$

Pregroup Derivation

$$\frac{Google}{NP} \quad \frac{bought}{NP^r \cdot S \cdot NP^l} \quad \frac{Microsoft}{NP}$$

Pregroup Derivation

$$
\dfrac{\dfrac{Google}{NP} \qquad \dfrac{\dfrac{bought}{NP^r \cdot S \cdot NP^l} \qquad \dfrac{Microsoft}{NP}}{NP^r \cdot S}}{}
$$

Pregroup Derivation

$$\dfrac{\dfrac{\mathit{Google}}{NP} \qquad \dfrac{\dfrac{\mathit{bought}}{NP^r \cdot S \cdot NP^l} \qquad \dfrac{\mathit{Microsoft}}{NP}}{NP^r \cdot S}}{S}$$

Various semantics spaces

- Lexical items of various grammatical types live in different 'spaces'.

$$\frac{man}{NP} \quad \frac{bites}{NP^r \cdot S \cdot NP^l} \quad \frac{dog}{NP}$$

$$\mathbf{N} \qquad \mathbf{N} \otimes \mathbf{S} \otimes \mathbf{N} \qquad \mathbf{N}$$

- Representations can be vectors or matrices.
- Basic types like nouns are vectors with components equal to TF*IDF values.
- Composition involves point-wise multiplication.

The sentence space

- What is the sentence space?
- Truth-theoretic interpretation: sentence space has two dimensions, **True** and **False**.
- Distributional interpretation: a point in the distributional space used for verbs. But what does this really mean (in particular in the case of complex sentences)??

Truth in a 2-dimensional space

dog chases cat

	⟨fluffy,T,fluffy⟩	⟨fluffy,F,fluffy⟩	⟨fluffy,T,fast⟩	⟨fluffy,F,fast⟩	⟨fluffy,T,juice⟩	⟨fluffy,F,juice⟩	⟨tasty,T,juice⟩ . . .
$\overrightarrow{chases}$	0.8	0.2	0.75	0.25	0.2	0.8	0.1
dog,cat	0.8,0.9	0.8,0.9	0.8,0.6	0.8,0.6	0.8,0.0	0.8,0.0	0.1,0.0

$$\overrightarrow{dog\ chases\ cat}_{\mathbf{T}} =$$
$$0.8 \,.\, 0.8 \,.\, 0.9 + 0.75 \,.\, 0.8 \,.\, 0.6 + 0.2 \,.\, 0.8 \,.\, 0.0 + 0.1 \,.\, 0.1 \,.\, 0.0 + \ldots$$

Sentence meaning in a multi-dimensional space

dog chases cat

	⟨fluffy,fluffy⟩	⟨fluffy,fast⟩	⟨fluffy,juice⟩	⟨tasty,juice⟩	⟨tasty,buy⟩	⟨buy,fruit⟩	⟨fruit,fruit⟩ . . .
$\overrightarrow{chases}$	0.8	0.75	0.2	0.1	0.2	0.2	0.0
dog,cat	0.8,0.9	0.8,0.6	0.8,0.0	0.1,0.0	0.1,0.5	0.5,0.0	0.0,0.0
$\overrightarrow{dog\ chases\ cat}$	0.576	0.36	0.0	0.0	0.01	0.0	0.0

Evaluation

- Evaluation against the phrase similarity task of Mitchell & Lapata (2010).
- Evaluation against a dataset of small sentences (e.g. *the table showed the results*).
- The pregroup grammar model outperforms simple pointwise methods on the sentence dataset.

Outline

Sense disambiguation: Erk & Padó (2008)

- Disambiguating *river bank*:
  1. COMPOUND-LEFT: river COMPOUND-RIGHT: bank
  2. Calculate centroid of word vectors which have COMPOUND-LEFT: river as context (average over *access, basin, boat*, etc)
  3. Compose (multiply) centroid with *bank* vector.

Disambiguating *bank*

bank

COMP^{-}:(compound)robber
COMP:(compound)savings
COMP^{-}:(compound)robbery
COMP:(of)Thames
COMP:(of)rhine
COOR:ditch
VERB:(ARG2)rob
COMP:(compound)sperm
COMP^{-}:(compound)account
COMP^{-}:(compound)Thai
COMP:(compound)Habib
COMP:(of)river
COMP:(of)River
COMP^{-}:(compound)Berhad
COMP:(compound)Deutsche
COMP:(of)Nile
COMP^{-}:(compound)teller
COMP:(compound)HSBC
COMP^{-}:(compound)holiday
COMP^{-}:(compound)Fargo

national bank

VERB:(ARG1)charge
COOR:strip
COOR:bed
COMP:(in)Philippines
VERB:(ARG2)rob
COOR:firm
VERB:(ARG2)burst
COMP^{-}:(on)section
VERB:(ARG1)borrow
COMP^{-}:(by)place_rel_
VERB:(ARG1)finance
COMP^{-}:(in)money
COOR:account
COMP^{-}:(of)failure
COMP^{-}:(from)money
VERB:(ARG2)bank
VERB:(ARG1)lower
COMP:(in)Hong_Kong
VERB:(ARG1)offset
COMP^{-}:(compound)Ltd

river bank

COMP:(of)stream
COMP(poss):river
COMP:(of)creek
COMP:(of)st
COMP:(of)canal
VERB:(ARG1)lend
COMP:(of)reservoir
COMP:(of)lake
COMP:(compound)river
COMP:(at)mouth
COMP^{-}:(on)village
COMP^{-}:(on)area
COMP:(of)Nile
COMP^{-}:(on)lie
COMP:(about)kilometer
VERB:(ARG2)erode
COMP^{-}:(on)situate
COOR:turn
COMP^{-}:(on)city
COMP:(of)channel

Outline

1. Compositionality
2. Composing distributions
   - Pointwise models
   - Lexical function model
   - Pregroup model
3. Sense disambiguation
4. Issues
   - Beyond intersection
   - Logical operators
   - Meaning (again!)
5. Conclusion

Beyond intersection

- What about non-intersective composition? (*fake, small, alleged*...)
- Even the semantics of intersective phrases is more than the intersection of their parts.

Is intersection enough?

A big city: just a city which is big?
See *loud, underground, advertisement, crowd, Phantom of the Opera*...

Adjective types, Partee (1995)

- **Intersective:** carnivorous mammal
  $||\text{carnivorous mammal}|| = ||\text{carnivorous}|| \cap ||\text{mammal}|$
- **Subsective:** skilful surgeon
  $||\text{skilful surgeon}|| \subseteq ||\text{surgeon}||$
- **Non-subsective:** former senator
  $||\text{former senator}|| \neq ||\text{former}|| \cap ||\text{senator}||$
  $||\text{former senator}|| \not\subseteq ||\text{senator}||$

Modelling classes of adjectives

- Boleda et al (2013).
- Compare composition functions on the three categories of adjectives.
- The lexical function model outperforms other methods.
- All methods perform just as well on the different categories.

What should we compose?

> *one has the common intuition that there is a perceived difference between [...] "Indian elephant" and "friendly elephant". [...] an Indian elephant is one of a recognized variety of elephants, and their properties are not simply those of being an elephant, and being from India, but something more (such as disposition, size of ears, etc. etc.) – it's a (sub)species. In this sense, "Indian elephant" differs from "friendly elephant" because a friendly elephant is no more than an elephant that is friendly, and that's it.*
>
> *Carlson (2010)*

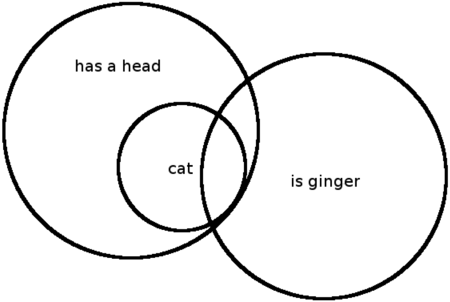
- What is the best representation for *Indian elephant*? The phrase or the composed form? Or both? (But how to do both??)

Logical operators

- Treatment of logical operators is unclear.
- In formal semantics, a quantifier ‘counts’ over the elements of a set.
  $Q(x)[rstr(x) \wedge scp(x)]$
  $\exists(x)[cat'(x) \wedge run'(x)]$
- No set in distributional semantics...

Quantifier entailment: Baroni et al (2012)

- Quantifiers: *no, few, some, many, most, all, more than two...*
- A stronghold of formal semantics: for sure, you need sets to do quantification...



Quantifier entailment: Baroni et al (2012)

- Study of *all, both, each, either, every, few, many, most, much, no, several, some*.
- Learn quantifier entailment by example: observe phrases such as *all cats/some cats* in a corpus, and train an SVM classifier.
- Classify previously unseen quantifier pairs.
- Results: up to 77% precision in detecting entailment. A surprising result.

The meaning of the sentence

- In formal semantics, meaning is denotational, compositional and truth-theoretic.
- *Kim sleeps* is true iff Kim is in the set of sleeping things: there is a systematic, compositional relation between the words in the sentence and the sets in the corresponding model.
- But distributions are more about intension than extension, so should we talk of denotation and truth?

Intension vs extension

- Extension (denotation):
  the things in the world that a word refers to.
- Intension:
  - *Morning star* vs. *Evening star* (the planet Venus);
  - the properties of a word (being visible in the morning for the Morning Star, in the evening for the Evening Star).
- DS is intensional in that it models things that are said about things (properties?), but not the things themselves.

Do distributions model meaning?

- A model of word meaning:
  - Cats are robots from Mars that chase mice.
  - Dogs are robots from Mars that chase cats.
  - Trees are 3D holograms from Jupiter.
- A similarity-based evaluation of this model would find that cats and dogs are very similar, but both are much less similar to trees.
- A good model of language?

Do distributions model meaning?

- A theory of meaning has to say how language relates to the world. For instance, model-theoretic semantics says that the meaning of *cat* is the set of all cats in a world.
- In distributionalism, meaning is the way we use words to talk *about* the world. No metaphysical assumptions.
- So if we use the words ‘robots from Mars’ to talk about cats, all is fine (see whales and fish).
- Not quite... (stay tuned: next week, ‘Formal Distributional Semantics’)

Outline

Conclusion

- We need a way to integrate lexical and compositional semantics.
- General feeling is that the composition of distributions should produce another distribution which expresses the meaning of a phrase/sentence.
- How to do this is only clear for certain constructions.
- What is the distribution of a sentence?
- How does this relate to meaning?