Distributional semantics for linguists

Lecture 3a: Distributional semantics and composition

Aurelie Herbelot

Universität Potsdam
Department Linguistik

ESSLLI 2012

Outline

Overview

- Composing distributions: the motivation. How to get from single words to phrases and sentences?
- Some compositional distributional models.
- Unanswered questions.

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?
- Also, 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

- Mitchell and Lapata (2010): word-based model, task-evaluated.
- Baroni and Zamparelli (2010): word-based, evaluated against phrasal distributions.
- Coecke, Sadrzadeh and Clark (2011): CCG-based model, task-evaluated.

Outline

The model

- 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.

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:
  $\exists x[\text{cat}'(x) \wedge \text{black}'(x)]$
  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)??

Addition

- Addition is not intersective: the whole meaning of both $\vec{u}$ and $\vec{v}$ are included in the resulting phrase.
- 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

Multiplication

- Multiplication is intersective.
- But it is commutative in a word-based model:
  $\overrightarrow{\text{The cat chases the mouse}} = \overrightarrow{\text{The mouse chases the cat}}$.

Outline

Overview

- Word-based model for adjective-noun composition.
- Composition is the multiplication of vectors/matrices learned from 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.
- There is no single composition operation for adjectives. Each adjective acts on nouns in a different way.

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
  $||\text{former senator}|| \neq ||\text{former}|| \cap ||\text{senator}||$
  $||\text{former senator}|| \not\subseteq ||\text{senator}||$

System

- For each adjective, a matrix is learned from actual AN phrases using partial least squares regression.
- Test by measuring distance between a given adjective-noun combination and the corresponding phrasal distribution.

Outline

Overview

- 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

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

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

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

Categorial Grammar Derivation

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

Categorial Grammar Derivation

$$
\begin{array}{ccc}
Google & bought & Microsoft \\
\hline
NP & NP\backslash S/NP & NP \\
 & \multicolumn{2}{c}{\overline{\hspace{12em}}>} \\
 & \multicolumn{2}{c}{NP\backslash S}
\end{array}
$$

Categorial Grammar Derivation

$$\dfrac{\dfrac{\textit{Google}}{NP} \qquad \dfrac{\dfrac{\textit{bought}}{NP\backslash S/NP} \qquad \dfrac{\textit{Microsoft}}{NP}}{NP\backslash S}{>}}{S}{<}$$

Pregroup Derivation

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

Pregroup Derivation

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

Pregroup Derivation

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

Various semantics spaces

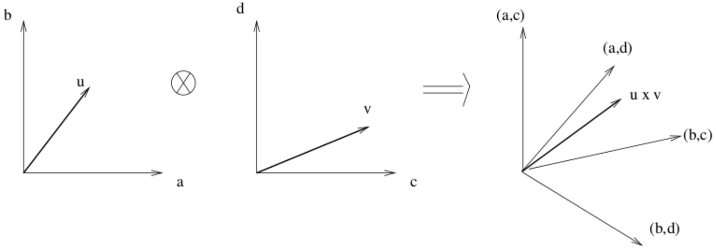
- Lexical items of various grammatical types live in different 'spaces'.

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

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

- Representations can be vectors or matrices.
  e.g. a transitive verb may be a matrix represented in a tensor product space $\mathbf{N} \otimes \mathbf{S} \otimes \mathbf{N}$.
- Basic types like nouns are vectors with components equal to TF*IDF values.
- Composition involves point-wise multiplication.

The tensor product



$$(u \otimes v)_{(a,d)} = u_a.v_d$$

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	

Outline

The meaning of the sentence

- In formal semantics, meaning is denotational and truth-theoretic.
- *Kim sleeps* is true iff Kim is in the set of sleeping things.
- Distributions are more about intension than extension, so should we talk of truth?
- If not, what should the meaning of a sentence be?

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*...

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.
  $Q(x)[\mathit{rstr}(x) \wedge \mathit{scp}(x)]$
  $\exists(x)[\mathrm{cat}'(x) \wedge \mathrm{run}'(x)]$
- No set in 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?