Distributional semantics for linguists
Lecture 3a: Distributional semantics and composition

Aurelie Herbelot
Universität Potsdam
Department Linguistik

ESSLLI 2012
Outline

1. Overview
2. Composing distributions: motivation
6. Issues
7. Conclusion
Overview

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

1. Overview
2. Composing distributions: motivation
6. Issues
7. Conclusion
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

- Baroni and Zamparelli (2010): word-based, evaluated against phrasal distributions.
Outline

1. Overview
2. Composing distributions: motivation
6. Issues
7. Conclusion
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 $p_i = \vec{u}_i \cdot \vec{v}_i$
  - tensor product $p_{ij} = \vec{u}_i \cdot \vec{v}_j$
  - circular convolution $p_{ij} = \sigma_j \vec{u}_j \cdot \vec{v}_{i-j}$
  - ... etc
- Task-based evaluation: similarity ratings. Multiplication is best measure.
### Example

<p>| early_j | africa::9.75873 | african::6.87337 | aftermath::3.40748 | afternoon::42.2096 | afterwards::7.46585 | again::9.00563 | age::15.6464 | aged::5.99896 | agencies::4.91747 | agency::7.28471 | agent::4.63014 | agents::4.21793 | ages::45.003 | ago::18.8909 | agree::5.05183 | agreed::6.36066 | agreement::7.64836 | agricultural::11.3745 |
| age_n   | africa::3.56225 | african::1.88733 | aftermath::1.37812 | afternoon::1.9041 | afterwards::3.86807 | again::2.78339 | age::0       | aged::24.6173 | agencies::1.57129 | agency::3.13776 | agent::2.24935 | agents::1.68319 | ages::0       | ago::19.2306 | agree::3.67157 | agreed::2.61272 | agreement::0.912126 | agricultural::2.66057 |
| early_j age_n | africa::34.76303 | african::12.97231 | aftermath::4.69591 | afternoon::80.3712 | afterwards::28.87843 | again::25.06618 | age::0       | aged::147.67819 | agencies::7.72677 | agency::22.85767 | agent::10.41480 | agents::7.09957 | ages::0       | ago::363.2833 | agree::18.54814 | agreed::16.61862 | agreement::6.976268 | agricultural::30.26265 |</p>
<table>
<thead>
<tr>
<th>Multiplication</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990s</td>
<td>talent</td>
</tr>
<tr>
<td>1980s</td>
<td>interested</td>
</tr>
<tr>
<td>1970s</td>
<td>showed</td>
</tr>
<tr>
<td>20th</td>
<td>learned</td>
</tr>
<tr>
<td>1960s</td>
<td>piano</td>
</tr>
<tr>
<td>Childhood</td>
<td>studying</td>
</tr>
<tr>
<td>1950s</td>
<td>exposed</td>
</tr>
<tr>
<td>Age</td>
<td>ages</td>
</tr>
<tr>
<td>1940s</td>
<td>parents</td>
</tr>
<tr>
<td>1920s</td>
<td>encouraged</td>
</tr>
<tr>
<td>1930s</td>
<td>singing</td>
</tr>
<tr>
<td>19th</td>
<td>educated</td>
</tr>
<tr>
<td>Late</td>
<td>interest</td>
</tr>
<tr>
<td>Century</td>
<td>uncle</td>
</tr>
<tr>
<td>Morning</td>
<td>violin</td>
</tr>
<tr>
<td>Stages</td>
<td>baronet</td>
</tr>
<tr>
<td>Settlers</td>
<td>eldest</td>
</tr>
<tr>
<td>Warning</td>
<td>raised</td>
</tr>
</tbody>
</table>
How do we interpret $f(\vec{u}, \vec{v})$ linguistically?

Intersection in formal semantics has a clear interpretation:
\[ \exists x [\text{cat}'(x) \land \text{black}'(x)] \]
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 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 is intersective.

But it is commutative in a word-based model:

\[
\text{The cat chases the mouse} = \text{The mouse chases the cat}.
\]
Outline

1. Overview

2. Composing distributions: motivation


6. Issues

7. Conclusion
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.
**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

1. Overview
2. Composing distributions: motivation
6. Issues
7. Conclusion
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
  \[ a^l \cdot a \rightarrow 1, \quad a \cdot a^r \rightarrow 1 \]

- The monoid is the set of grammatical types ($NP$, $NP^r$, $NP^l$, $NP^{rr}$, $NP^{ll}$, $S$, $PP$, ...) 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

\[
\begin{align*}
\text{Google} & \quad \text{bought} & \quad \text{Microsoft} \\
NP & \quad NP\backslash S/NP & \quad NP
\end{align*}
\]
Categorial Grammar Derivation

\[
\begin{array}{c}
\text{Google} \\
NP
\end{array}
\quad \begin{array}{c}
bought \\
NP \backslash S / NP
\end{array}
\quad \begin{array}{c}
\text{Microsoft} \\
NP
\end{array}
\quad \begin{array}{c}
\rightarrow \\
NP \backslash S
\end{array}
\]
Categorial Grammar Derivation

\[
\begin{array}{c}
\text{Google} \\
\text{bought} \\
\text{Microsoft}
\end{array}
\quad
\begin{array}{c}
\text{NP} \\
\text{NP} \backslash S \slash NP \\
\text{NP}
\end{array}
\quad
\begin{array}{c}
\text{NP} \\
\text{NP} \backslash S
\end{array}
\quad
\begin{array}{c}
S
\end{array}
\]

\[
\frac{\text{NP}}{S}
\]
Pregroup Derivation

\[
\begin{array}{ccc}
\underline{Google} & \underline{bought} & \underline{Microsoft} \\
NP & NP^r \cdot S \cdot NP^l & NP \\
\end{array}
\]
Pregroup Derivation

\[
\text{Google} \quad \frac{\text{bought}}{NP^r \cdot S \cdot NP^l} \quad \text{Microsoft}
\]

\[
NP = NP^r \cdot S
\]
Pregroup Derivation

\[
\begin{align*}
\text{Google} & \quad \underline{NP} \\
\text{bought} & \quad \underline{NP^r \cdot S \cdot NP^l} \\
\text{Microsoft} & \quad \underline{NP} \\
\text{S} & \quad \underline{NP^r \cdot S}
\end{align*}
\]
Lexical items of various grammatical types live in different ‘spaces’.

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

\[(u \otimes v)_{(a,d)} = u_a \cdot v_d\]
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)??

Coecke et al (2010)

Herbelot, Aurelie (Universität Potsdam) Distributional semantics for linguists ESSLLI 2012 28 / 37
Truth in a 2-dimensional space

\[
\text{dog chases cat}
\]

<table>
<thead>
<tr>
<th></th>
<th>fluffy,T,fluffy</th>
<th>fluffy,F,fluffy</th>
<th>fluffy,T,fast</th>
<th>fluffy,F,fast</th>
<th>fluffy,T,juice</th>
<th>fluffy,F,juice</th>
<th>tasty,T,juice</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathbf{chases})</td>
<td>0.8</td>
<td>0.2</td>
<td>0.75</td>
<td>0.25</td>
<td>0.2</td>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td>(\mathbf{dog,cat})</td>
<td>0.8,0.9</td>
<td>0.8,0.9</td>
<td>0.8,0.6</td>
<td>0.8,0.6</td>
<td>0.8,0.0</td>
<td>0.8,0.0</td>
<td>0.1,0.0</td>
</tr>
</tbody>
</table>

\[
\text{dog chases cat}_T = 0.8 \cdot 0.8 \cdot 0.9 + 0.75 \cdot 0.8 \cdot 0.6 + 0.2 \cdot 0.8 \cdot 0.0 + 0.1 \cdot 0.1 \cdot 0.0 + \ldots
\]
Sentence meaning in a multi-dimensional space

\[ \text{dog chases cat} \]

<table>
<thead>
<tr>
<th></th>
<th>fluffy,fluffy</th>
<th>fluffy,fast</th>
<th>fluffy,juice</th>
<th>tasty,juice</th>
<th>tasty,buy</th>
<th>buy,fruit</th>
<th>fruit,fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>chases</strong></td>
<td>0.8</td>
<td>0.75</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>dog,cat</strong></td>
<td>0.8,0.9</td>
<td>0.8,0.6</td>
<td>0.8,0.0</td>
<td>0.1,0.0</td>
<td>0.1,0.5</td>
<td>0.5,0.0</td>
<td>0.0,0.0</td>
</tr>
<tr>
<td><strong>dog chases cat</strong></td>
<td>0.576</td>
<td>0.36</td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Outline

1. Overview
2. Composing distributions: motivation
6. Issues
7. Conclusion
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? \((\text{fake}, \text{small}, \text{alleged}...\)\)
- 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*...
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.
  \[ Q(x)[rstr(x) \land scp(x)] \]
  \[ \exists(x)[cat'(x) \land run'(x)] \]
- No set in distributional semantics...
Conclusion

Outline

1 Overview

2 Composing distributions: motivation

3 Mitchell and Lapata (2010)

4 Baroni and Zamparelli (2010)

5 Coecke et al (2010)

6 Issues

7 Conclusion
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?