Compositionality and Distributional Semantic Models
An introduction

Alessandro Lenci\textsuperscript{1}, Roberto Zamparelli\textsuperscript{2}

\textsuperscript{1}University of Pisa, Department of Linguistics
\textsuperscript{2}University of Trento, CIMeC

Copenhagen, August 16 2010
Outline

1. Distributional semantics
   - DSMs in a nutshell

2. Compositionality

3. Compositionality in DSMs
Distributional semantics

- **Distributional Semantic Models (DSMs)** aim at characterizing the meaning of linguistic expressions in terms of their distributional properties.

- DSMs all rely on some version of the **Distributional Hypothesis (DH)**; Harris 1954, Miller & Charles 1991:
  
  - at least certain aspects of the meaning of lexical expressions depend on their distributional properties in the linguistic contexts.
  
  - the degree of semantic similarity between two linguistic expressions $A$ and $B$ is a function of the similarity of the linguistic contexts in which $A$ and $B$ can appear.

- The format of **distributional representations** greatly varies depending on the specific aspects of meaning they are designed to model.
Weak and Strong DH
Lenci (2008)

Weak DH
A quantitative method for semantic analysis and lexical resource induction
- word meaning (*whatever this might be*) is reflected in linguistic distributions
- by inspecting a relevant number of distributional contexts, we may identify those aspects of meaning that are shared by words that have similar contextual distributions

Applications
- E-language modeling, lexicography, NLP
  - word sense disambiguation, ontology and thesauri learning, relation extraction, question answering, etc.
Strong DH

A cognitive hypothesis about the form and origin of semantic representations

- word distributions in context have a specific causal role in the formation of the semantic representation for that word
- the distributional properties of words in linguistic contexts explains human semantic behavior (e.g. judgments of semantic similarity)

applications  I-language modeling, concept modeling
- semantic priming, word learning, semantic deficits, etc.
Distributional Semantic Models (DSMs)

- Computational models that build **contextual semantic representations** from corpus data

- DSMs are models for **semantic representations**...
  - the semantic content is represented by a **vector**

  ... and for **the way semantic representations are built**
  - vectors are obtained through the statistical analysis of the linguistic contexts of a word

- **Alternative names for DSMs**
  - *corpus-based semantics*
  - *statistical semantics*
  - *geometrical models of meaning*
  - *vector semantics*
  - *word (semantic) space models*
A taxonomy of DSMs
Baroni & Lenci (in press), Baroni & Lenci (in preparation)

- **Geometrical DSMs**
  - Latent Semantic Analysis (LSA; Landauer & Dumais 1997)
  - Hyperspace Analogue to Language (HAL; Lund & Burgess 1996)
  - Dependency Vectors (DV; Padó & Lapata 2007)
  - Distributional Memory (DM; Baroni & Lenci in press)
  - ... among many others variations

- **Probabilistic DSMs**
  - Topic Models (Griffiths et al. 2007)
Distributional vectors

- **Count** how many times each target word occurs in a certain context
- **Build vectors** out of (a function of) these context occurrence counts

**Hypothesis**

- Semantically similar words will have *similar vectors*
- DH is the “bridging assumption” that turns distributional similarity into semantic similarity
A general definition of DSMs

- DSMs are tuples \(< T, C, D, W, M, d, S >\)
  - **T** target elements, i.e., the words for which the DSM provides a contextual semantic representation
  - **C** contexts, with(in) which \( T \) (co)occur
    - e.g., documents, neighbor words, syntactically related words, etc.
  - **D** domain, within which to consider the contexts
    - e.g., a corpus, a particular set of documents, the WWW, etc.
  - **W** context weighting scheme
    - e.g., frequency, log-frequency, association measures, etc.
  - **M** distributional matrix, \( T \times C \)
  - **d** dimensionality reduction function, \( d : M \rightarrow M' \)
    - e.g. Singular Value Decomposition (SVD), etc.
  - **S** distance measure, between the vectors in \( M' \)
    - e.g. cosine, etc.
Contextual representations as distributional vectors

distributional matrix  =  targets \times contexts

<table>
<thead>
<tr>
<th>targets</th>
<th>leash</th>
<th>walk</th>
<th>run</th>
<th>owner</th>
<th>leg</th>
<th>bark</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>cat</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>lion</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>light</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bark</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Semantic similarity as angle between vectors

cat (1,5)
dog (1,4)
car (4,0)
DSMs interpret semantic similarity as a **quantitative notion**

if $\vec{a}$ is closer to $\vec{b}$ in the distributional vector space, than $a$ is more semantically similar to $b$.

<table>
<thead>
<tr>
<th>rhino</th>
<th>fall</th>
<th>rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>woodpecker</td>
<td>rise</td>
<td>lava</td>
</tr>
<tr>
<td>rhinoceros</td>
<td>increase</td>
<td>sand</td>
</tr>
<tr>
<td>swan</td>
<td>fluctuation</td>
<td>boulder</td>
</tr>
<tr>
<td>whale</td>
<td>drop</td>
<td>ice</td>
</tr>
<tr>
<td>ivory</td>
<td>decrease</td>
<td>jazz</td>
</tr>
<tr>
<td>plover</td>
<td>reduction</td>
<td>slab</td>
</tr>
<tr>
<td>elephant</td>
<td>logarithm</td>
<td>cliff</td>
</tr>
<tr>
<td>bear</td>
<td>decline</td>
<td>pop</td>
</tr>
<tr>
<td>satin</td>
<td>cut</td>
<td>basalt</td>
</tr>
<tr>
<td>sweatshirt</td>
<td>hike</td>
<td>crevice</td>
</tr>
</tbody>
</table>
Outline

1. Distributional semantics
   - DSMs in a nutshell

2. Compositionality

3. Compositionality in DSMs
Fodor & Lepore (2002:1)

“Compositionality is the property that a system of representation has when
(i) it contains both primitive symbols and symbols that are syntactically and semantically complex;
(ii) the latter inherit their semantic properties from the former”
Arguments for compositionality

- Compositionality is claimed to be a necessary condition for semantic representations to address two key properties of natural language
  - **Productivity** - the ability to produce an infinite number of distinct meaningful expressions
    - *A cat chases a mouse, A big cat chases a brown mouse, A big black cat chases a little brown mouse, A big black cat chases a little brown mouse that chases a young red cat, etc.*
    - cf. “this would be impossible were we not able to distinguish parts in the thoughts corresponding to the parts of a sentence” (Frege 1923)
  - **Sistematicity** - the ability to produce/understand some linguistic expressions is intrinsically related to the ability to produce/understand some other expressions
    - cf. if we understand *A cat chases a mouse* we also understand *A mouse chases a cat* (and we also understand that they mean very different things...)
The principle of compositionality

The meaning of a complex expression is a function of the meanings of its parts and of their syntactic mode of combination.

- The ingredients of compositionality (Partee 1984)
  - a theory of lexical meanings - assigns meanings to the smallest part (e.g. words)
  - a theory of syntactic structures - determines the relevant part-whole structure of each complex expression
  - a theory of semantic composition - determines the combinatorial semantic operations, i.e. the functions that compose the meanings
Compositionality in formal semantics

- In denotational semantics we have a clear assumption about the interpretation of complex expressions, i.e. sentences
  - sentences denote truth-values \((t)\) or propositions \((<w,t>)\)
- The denotation of the component expressions is their contributions to the computation of the sentence denotation

<table>
<thead>
<tr>
<th>word</th>
<th>type</th>
<th>logical form</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>e</td>
<td>Tom</td>
<td>Tom, the cat</td>
</tr>
<tr>
<td>chases</td>
<td>&lt;e, e, t&gt;</td>
<td>(\lambda y \lambda x.\text{chase}(x,y))</td>
<td>({&lt;x,y&gt;</td>
</tr>
<tr>
<td>Jerry</td>
<td>e</td>
<td>Jerry</td>
<td>Jerry, the mouse</td>
</tr>
</tbody>
</table>

- The key operation for semantic composition is type-driven functional application (possibly integrated with other types of operations; cf. Pustejovsky 1995, et al.)
  - \(\lambda y \lambda x.\text{chase}(x,y)(\text{Jerry}) = \lambda x.\text{chase}(x,\text{Jerry}) \Rightarrow \{<x, \text{Jerry} > | x \text{ chases Jerry the mouse}\}\)
Outline

1. Distributional semantics
   - DSMs in a nutshell

2. Compositionality

3. Compositionality in DSMs
DSMs and formal semantics

- Current DSMs are typically geared towards representing the meaning of *words in isolation*.
  - Typical semantic tasks addressed by DSMs involve similarity between words (attributational similarity) or between word pairs (relational similarity) (Turney 2007, Baroni & Lenci in press).
  - No attempt to build DSM for larger constituents, nor to provide an interface between DS lexical representations and non-DS models of constituent meaning.

- DSM (at least in NLP) have never tried to characterize the meaning of *function words or morphemes*. Typically, functional elements are not even included in the context.
  - In contrast, most work in formal semantics has been devoted to functional elements (e.g. quantifiers, Wh-elements), constructions (e.g. bare plurals) or general classes of content words (e.g. mass vs. count nouns, verb aktionsart)

- DSMs have at their base *frequency of cooccurrence* in a corpus. Formal semantics rests on the notion of *truth* in a model. There might be a relation, but corpora are easily inconsistent.
DSMs for larger-than-word fragments

Generalized DS

In principle it is perfectly possible and perhaps sensible to extract DS for multi-word expressions: not just “run” and “race”, but also “run the race”.

- After all, discovering that “Hi!”, “What’s up?” and “Good morning”, despite their syntactic differences, appear in similar context does provide valuable information about the usage of such expressions.
- Other cases where extracting multi-word expressions seems desirable are those where compositionality fails, e.g. idioms.
  - John {died / kicked the bucket}
  - Marc {showered / took a shower}
However, as one moves from idioms and collocations to more general multi-word constituents (VPs, NPs, S), a huge problem of data sparseness develops:

**Google results:**
- “girl” 619,000,000
- “kissed” 23,400,000
- “boy” 669,000,000
- “kissed a boy” 7,180,000
- “a girl kissed a boy” 6

Evidently, if we want a DS for strings as large as whole constituents, we need a way to build them *compositionally*. 
DS for constituents?

But, is there a real need to combine DSM with the compositionality-based formal semantic approach?

Why should we care?

Two points in favor (more, please!)
- Generality
- Lexical Ambiguity
Generality

- At the lexical level, DSM can model psychological notion like priming, semantic similarity, etc. (Miller & Charles 1991, Griffiths et al. 2007).
- Compositionality is a crucial testbed to evaluate whether DMSs can offer a plausible model for (some aspects of) semantic representations in general.
  - for instance, Fodor et al. have used the failure to account for compositionality as an argument against particular types of representations (i.e. connectionist models, concepts as prototypes, etc.)
  - “since mental representations are de facto compositional, we can reject out of hand any theory that says that concepts(/word meanings) are Xs unless Xs are the sorts of things for which compositionality holds” (Fodor & Lepore 2002: 3)
Lexical Ambiguity

Most (frequent) words are syntactically and/or semantically ambiguous

- $\text{run}_1$ a maraton
- $\text{run}_2$ an experiment
- $\text{run}_3$ a lab
- The horse $\text{ran}_4$.
- The colors $\text{ran}_5$.

Widespread tacit assumption in formal semantics: words are semantically (as well as syntactically) disambiguated before being translated into logical forms:

- $\llbracket \text{Bolt ran} \rrbracket = \text{run'}_4(\text{bolt})$
- $\llbracket \text{the color ran} \rrbracket = \text{run'}_5(\text{the\_color})$
Lexical Ambiguity

But this disambiguation process is unrealistic, since it requires semantic knowledge about what are the elements that are being combined (co-composition).

Possible answer: semantics generates all possible compositions of all (type-compatible) meanings for the words it needs to combine.

- Prediction: large number of spurious meanings.
- Not psychologically tenable.
DSM can be used to disambiguation words. They could thus provide a first level of ambiguity reduction that “feeds” further semantic compositions. Cf. Erk & Pado 2008.

\[ \text{Verb(Noun)} \Rightarrow \text{Verb-as-a-predicate-that-takes-Bs(Noun-as-an-object-that-is-taken-by-As)} \]

However, Erk & Pado do not actually provide a single DS for the “Verb Noun” constituent.
1. What are the **semantic operations** that drive compositionality in DSMs?

2. What is the **interpretation** to be assigned to complex expressions (e.g. phrases, sentences, etc.) in DSMs? What is the relation between the vectors generated by DSM and the interpretation function $[[ ]]$?

3. How to represent the meaning of words in context?

4. Which cases need to be fully **recursive**?

5. How to account for the dependence of compositional meaning on **syntactic structure**?

6. How to test compositional representations in DSMs?
Semantic compositional operations in DSMs

Composition as vector combination

The standard approach is to define the distributional “meaning” of a phrase as the combined vector (e.g., the sum of their vectors) built with the vectors of the words in the phrase

Types of vector composition
Mitchell & Lapata (2010)

- **Simple vector sum** (Landauer & Dumais 1997)
  \[ \overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} \]
  \[ \overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat} \]

- **Contexts-sensitive vector sum** (Kintsch 2001)
  \[ \overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} + \sum \overrightarrow{n} \]
  \( n \) are the \( n \)-top nearest neighbors of the predicate which are also neighbors of the argument.
  \[ \overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat} + ( \overrightarrow{hunt} + \overrightarrow{prey} + \ldots + \overrightarrow{capture} ) \]
  Kintsch captures effects of context-sensitivity in predication (e.g. disambiguation, co-composition, metaphorical interpretation, etc.)

- **Vector pairwise multiplication** (Mitchell & Lapata 2010)
  \[ \overrightarrow{p} = \overrightarrow{a} \cdot \overrightarrow{b} \]
  \[ \overrightarrow{chase\ cat} = \overrightarrow{chase} \cdot \overrightarrow{cat} \]
The interpretation of complex expressions in DSMs

- In denotational semantics, the outcome of compositional operations has a straightforward interpretation
  \[ \lambda y \lambda x. \text{chase}(x,y)(\text{Jerry}) = \lambda x. \text{chase}(x,\text{Jerry}) \Rightarrow \text{the set of individuals that chase Jerry the mouse} \]

- In DSMs we have a clear idea of what word distributional vectors represent
  - the most prototypical (i.e. statistically significant) linguistic contexts in which a word occurs

- We don’t have clear intuitions of what a composed vector stands for semantically
  - e.g.
  - \( \text{chase cat} \) is something in between \( \text{chase} \) and \( \text{cat} \), but the meaning of \( \text{chase cat} \) is not something in between the meaning of \( \text{chase} \) and the meaning of \( \text{cat} \)
Addressing compositionality entails accounting how the meaning of complex expressions depends on their syntactic structure (i.e. going beyond “bag of words” approaches)

- e.g. *A cat chases a mouse ≠ A mouse chases a cat*

Simple operations of vector compositions (e.g. sum, pairwise multiplication) are commutative

- i.e., *A cat chases a mouse = A mouse chases a cat*

Some explored solutions:

- use non-symmetric vector composition operations (Mitchell & Lapata 2010)
- make vector construction sensitive to word order (Jones & Mewhort 2007, Sahlgren *et al.* 2008)
- use syntax-based vectors (Erk & Padó 2008)
Compositionality and contextual effects

- Compositionality could be a very simple process, but it is complicated by the behaviour of lexical items in context.
  - non-intersectivity
    - *skillful politician* vs. *fast turtle* vs. *stone lion*
  - coercion
    - *enjoy a book* vs. *enjoy an ice cream*


“It would seem that part of knowing the meaning of a word should have to involve knowing how the basic meaning(s) could be stretched, shrunk, or otherwise revised in various ways when necessary; since the possible revisions are probably not finitely specifiable, such a conception of meaning would take us well beyond the normal conception of the lexicon as a finite list of finite specifications of idiosyncratic information about a particular lexical items”
When words are composed, they tend to affect each other’s meanings

- *The horse runs* vs. *The water runs*
- “The horse horse-like runs”
- cf. an instance of context-sensitive interpretation of lexical items

Erk & Padó (2008)

- *run* vector in the context of *horse* is a (multiplicative or additive) combination of the *run* vector and a prototype vector that represents the typical verbs *horse* is a subject of

  - \[ \text{run-in-the-context-of-horse} = \overrightarrow{\text{run}} \cdot (\overrightarrow{\text{gallop}} + \overrightarrow{\text{trot}} + \ldots) \]

- *horse* vector in the context of *run* is a (multiplicative or additive) combination of the *horse* vector and a prototype vector that represents the typical subjects of *run*

  - \[ \text{horse-in-the-context-of-run} = \overrightarrow{\text{horse}} \cdot (\overrightarrow{\text{car}} + \overrightarrow{\text{lion}} + \ldots) \]
Very little work has been done outside the domain of verb-argument composition, which is not recursive. *Modification* is recursive:

```
  house
  [red house]
  [large [red house]]
  [ugly [large [red house]]]
```

But if “red” is to combine with “house” in the same way “large” combines with “red house”, the type of DS produced by “house” must be the same as that produced by “red house”.

Applications: inconsistency (“even odd number”), semantic similarity (“black rose” – “rare”)
Testing compositionality in DSMs

- In denotational semantics, the outcome of compositional operations can be tested by checking truth-conditions, inferences, etc.

- In DSMs, the evaluation for semantic representations is represented by semantic tasks based on semantic similarity between words or word pairs (e.g., TOEFL, SAT analogy test, etc.)
  - this operational evaluation is crucial, because distributional vectors can not be directly “inspected”

- To address compositionality in DSMs we need to define relevant (and plausible...) semantic tasks with which to evaluate the composed vector representations operationally
  - paraphrasing (Erk & Padó 2008)
  - “phrase similarity” (Mitchell & Lapata 2010)
  - ...