Will Distributional Semantics Ever Become Semantic?

Alessandro Lenci

University of Pisa

7th International Global WordNet Conference
January 28, 2014 - Tartu
Distributional semantics is predicated on the assumption that linguistic units with certain semantic similarities also share certain similarities in the relevant environments.

If therefore relevant environments can be previously specified, it may be possible to group automatically all those linguistic units which occur in similarly definable environments, and it is assumed that these automatically produced groupings will be of semantic interest.

**What is Distributional Semantics?**

*Distributional semantics* is predicated on the assumption that *linguistic units with certain semantic similarities also share certain similarities in the relevant environments.*

If therefore relevant environments can be previously specified, it may be possible to group automatically all those linguistic units which occur in similarly definable environments, and it is assumed that these automatically produced groupings will be of *semantic interest.*

To be relevant [linguistic] elements must be set up on a **distributional basis**: $x$ and $y$ are included in the same element $A$ if the distribution of $x$ relative to the other elements $B$, $C$, etc. is in some sense the same as the distribution of $y$.

(Harris 1951: 7)
If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution.

(Harris 1954: 156)
A linguist defines the distribution of a word as the list of contexts into which the word can be substituted; the distributional similarity of two words is thus the extent to which they can be substituted into the same contexts. [...] Several psychologist have invented or adapted variations on this distributional theme as an empirical method for investigating semantic similarities.

(Miller 1967: 572-573)
The contextual representation of a word is knowledge of how that word is used. […] That is to say, a word’s contextual representation […] is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. […] Two words are semantically similar to the extent that their contextual representations are similar.

(Miller and Charles 1991: 5)
The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- Distributional semantic models (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity
The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- **Distributional semantic models** (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity
The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- **Distributional semantic models** (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity
The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- **Distributional semantic models** (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity
The Theoretical Foundations of Distributional Semantics

The Distributional Hypothesis

Lexemes with similar distributional properties have similar meanings

- **Distributional semantic models** (DSMs) are computational methods that turn the Distributional Hypothesis into an experimental framework for semantic analysis:
  - extract from corpora and count co-occurrences of lexical items with linguistic contexts
  - represent lexical items geometrically with distributional vectors built out of (a function of) their co-occurrence counts
  - measure semantic similarity with distributional vector similarity
... dig a [hole. The 
... to directly [drive the 
... celebrity status, [drove fast 
... but there [are police 
... world of [money, fast 
... to pet [the family's 
... and then [wanted a 
... murmur is [detectable. The 
... behaviour of [a domestic 
... have never [seen a 
... bank, children [playing with 
... sure you [encourage your 
... Truth, Lord: [yet the 
... vegetable material [and enzymes. 
... hubby once [ate the 
... were back [at the 
... go down [as the 
... heavy objects, [driving transit 
... of the [fast food 
... each of [the six 

car drove away] leaving behind ...
car wheel angle] 3. Force ...
cars and partied] with some ...
cars that chase] you. Each ...
cars and excitement] and, under ...
cat and dog.] who tended ...
cat to eat] the many ...
cat often eats] and drinks ...
cat playing with] a caught ...
cat eat so] little and ...
dogs and a] man leading. ... 
dog to play] appropriate chase ...
dogs eat of] the crumbs ...
Dogs also eat] fruit, berries ...
dog food and] asked for ...
van and drove] down to ...
van drove off.] As he ...
vans , wiring plugs] and talking ...
van being located] outside their ...
van wheels, and] also under ...
... to Distributional Vectors

\[
\begin{pmatrix}
\text{dog} & \text{drive} & \text{eat} & \text{fast} & \text{play} & \ldots & \text{the} & \text{wheel} \\
\text{car} & 0 & 3 & 0 & 2 & 0 & \vdots & 2 & 1 \\
\text{cat} & 1 & 0 & 3 & 0 & 1 & \vdots & 2 & 0 \\
\text{dog} & 0 & 0 & 3 & 0 & 2 & \vdots & 2 & 0 \\
\text{van} & 0 & 3 & 0 & 1 & 0 & \vdots & 3 & 1 \\
\end{pmatrix}
\]

coop-occurrence matrix
... to Distributional Vectors

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>drive</th>
<th>eat</th>
<th>fast</th>
<th>play</th>
<th>...</th>
<th>the</th>
<th>wheel</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>:</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>:</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>:</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>van</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>:</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

co-occurrence matrix
... to Distributional Vectors
Measuring Vector Similarity

Cosine

$$\frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$
Measuring Vector Similarity

Euclidean distance

\[ \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \]

Alessandro Lenci

GWC 2014 @ Tartu - January 28, 2014
The Distributional Hypothesis predicts that words with similar distributional vectors are semantically similar.

<table>
<thead>
<tr>
<th></th>
<th>car</th>
<th>cat</th>
<th>dog</th>
<th>van</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>0.33</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>0.60</td>
<td>0.94</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>van</td>
<td>0.92</td>
<td>0.50</td>
<td>0.76</td>
<td>1</td>
</tr>
</tbody>
</table>

cosine similarities
The Distributional Hypothesis predicts that words with similar distributional vectors are semantically similar.

<table>
<thead>
<tr>
<th></th>
<th>car</th>
<th>cat</th>
<th>dog</th>
<th>van</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>0.33</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>0.60</td>
<td>0.94</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>van</td>
<td>0.92</td>
<td>0.50</td>
<td>0.76</td>
<td>1</td>
</tr>
</tbody>
</table>

cosine similarities
Types of DSMs

- **Linguistic contexts**
  - text regions, linear (window-based) collocates, syntactic collocates, etc.

- **Context weighting**
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.

- **Distributional vector construction**
  - matrix models, topic models, Random Indexing, neural embeddings, etc.

- **Vector similarity measure**
  - cosine, euclidean distance, LIN measure, etc.
Types of DSMs

- Linguistic contexts
  - text regions, linear (window-based) collocates, syntactic collocates, etc.

- Context weighting
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.

- Distributional vector construction
  - matrix models, topic models, Random Indexing, neural embeddings, etc.

- Vector similarity measure
  - cosine, euclidean distance, LIN measure, etc.
Types of DSMs

- **Linguistic contexts**
  - text regions, linear (window-based) collocates, syntactic collocates, etc.

- **Context weighting**
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.

- **Distributional vector construction**
  - matrix models, topic models, Random Indexing, neural embeddings, etc.

- **Vector similarity measure**
  - cosine, euclidean distance, LIN measure, etc.
Types of DSMs

- **Linguistic contexts**
  - text regions, linear (window-based) collocates, syntactic collocates, etc.

- **Context weighting**
  - raw co-occurrence frequency, entropy, tf-idf, association measures, etc.

- **Distributional vector construction**
  - matrix models, topic models, Random Indexing, neural embeddings, etc.

- **Vector similarity measure**
  - cosine, euclidean distance, LIN measure, etc.
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity
- The Distributional Hypothesis is primarily a conjecture about word meaning
- Distributional semantics is based on a holistic and relational view of meaning
- Distributional semantics is based on a contextual and usage-based view of meaning
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts
- Distributional representations are quantitative, continuous, gradable and distributed
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about **semantic similarity**
- The Distributional Hypothesis is primarily a conjecture about **word meaning**
- Distributional semantics is based on a **holistic and relational** view of meaning
- Distributional semantics is based on a **contextual and usage-based** view of meaning
- Distributional semantics represent lexemes with **distributional vectors** recording their **frequency distribution** in linguistic contexts
- Distributional representations are **quantitative, continuous, gradable and distributed**
The Main Characters of Distributional Semantics

- The Distributional Hypothesis is primarily a conjecture about semantic similarity.
- The Distributional Hypothesis is primarily a conjecture about word meaning.
- Distributional semantics is based on a holistic and relational view of meaning.
- Distributional semantics is based on a contextual and usage-based view of meaning.
- Distributional semantics represent lexemes with distributional vectors recording their frequency distribution in linguistic contexts.
- Distributional representations are quantitative, continuous, gradable and distributed.
Weak and Strong Distributional Hypothesis

Lenci (2008)

Weak Distributional Hypothesis

An empirical method for semantic analysis

- 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 lexicography, ontology and thesauri learning and population, word sense disambiguation, relation extraction, question answering, Information Retrieval, etc.
Weak and Strong Distributional Hypothesis
Lenci (2008)

Weak Distributional Hypothesis

An empirical method for semantic analysis
- 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
lexicography, ontology and thesauri learning and population, word sense disambiguation, relation extraction, question answering, Information Retrieval, etc.
Weak and Strong Distributional Hypothesis
Lenci (2008)

Strong Distributional Hypothesis

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 is an explanatory factor of human semantic competence

applications: models of semantic memory (e.g. semantic priming, categorization, etc.), word learning, semantic processing, etc.
Weak and Strong Distributional Hypothesis

Lenci (2008)

Strong Distributional Hypothesis

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 is an explanatory factor of human semantic competence

Applications

models of semantic memory (e.g. semantic priming, categorization, etc.), word learning, semantic processing, etc.
The answer can be based on a particular “pre-conception” of meaning

As Wittgenstein says, ‘the meaning of words lies in their use.’
J. R. Firth (1951), “Modes of meaning”
Semantics with no treatment of truth conditions is not semantics
D. Lewis (1970), “General semantics”
To know the meaning of a sentence is to know its truth conditions
I. Heim and A. Kratzer (1998), Semantics in Generative Grammar

The answer can be based on the types of semantic facts that
distributional semantics is able to explain
The answer can be based on a particular “pre-conception” of meaning.

As Wittgenstein says, ‘the meaning of words lies in their use.’
J. R. Firth (1951), “Modes of meaning”

*Semantics with no treatment of truth conditions is not semantics*
D. Lewis (1970), “General semantics”

*To know the meaning of a sentence is to know its truth conditions*
I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*

The answer can be based on the types of semantic facts that distributional semantics is able to explain.
The answer can be based on a particular “pre-conception” of meaning:

As Wittgenstein says, ‘the meaning of words lies in their use.’
J. R. Firth (1951), “Modes of meaning”

Semantics with no treatment of truth conditions is not semantics
D. Lewis (1970), “General semantics”

To know the meaning of a sentence is to know its truth conditions
I. Heim and A. Kratzer (1998), *Semantics in Generative Grammar*

The answer can be based on the types of semantic facts that distributional semantics is able to explain.
How Semantic is Distributional Semantics?

Distributional Semantics and its Boundaries

Success stories

- Semantic similarity
  - synonymy, categorization, etc.
- Selectional preferences
  - semantic typing, co-composition, etc.
- Context-based semantic phenomena
  - sense-shifts, gradience, world knowledge integration, etc.
- Figurative language
  - analogy, metaphor, etc.
- Cognitive modeling
  - semantic priming, similarity judgements, thematic fit, etc.
Distributional Semantics and its Boundaries

Terrae incognitae

- **Function words**
  - negation, quantification, logical connectives, discourse particles, ecc.

- **Intensionality**
  - tense, aspect, modality, etc.

- **Reference and coreference**
  - indexicals, anaphora, etc.
Current challenges

- **Polysemy**
  - sense induction, regular polysemy, etc.
- **Compositionality**
  - adjectival modification, predicate-argument structures, etc.
- **Semantic relations**
  - hypernymy, antonymy, etc.
- **Inference**
  - lexical entailments, presuppositions, implicatures, etc.
Similar words differ for the type of relation holding between them
- _dog_ is very similar to both _animal_ and _cat_, but _animal_ is an _hyponym_ and _cat_ is a _coordinate_ (co-hyponym)

DSMs provide a quantitative correlate of semantic similarity (relatedness), but do not discriminate between different types of semantic relations
- cf. WordNet instead provides a “typed” semantic space
Similar words differ for the type of relation holding between them

- *dog* is very similar to both *animal* and *cat*, but *animal* is an hypernym and *cat* is a coordinate (co-hyponym)

DSMs provide a quantitative correlate of semantic similarity (relatedness), but do not discriminate between different types of semantic relations

- cf. *WordNet* instead provides a “typed” semantic space
Semantic Relations


- **synonymy**  *sofa* - *couch*
- **hyperonymy**  *dog* - *animal*
- **co-hyponymy**  *dog* - *cat*
- **antonymy**  *dead* - *alive*
- **meronymy**  *wheel* - *car*
### Distributional Neighbors from the BNC

#### dog (window size= 2)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>0.77</td>
</tr>
<tr>
<td>horse</td>
<td>0.67</td>
</tr>
<tr>
<td>fox</td>
<td>0.65</td>
</tr>
<tr>
<td>pet</td>
<td>0.63</td>
</tr>
<tr>
<td>rabbit</td>
<td>0.61</td>
</tr>
<tr>
<td>pig</td>
<td>0.57</td>
</tr>
<tr>
<td>animal</td>
<td>0.57</td>
</tr>
<tr>
<td>mongrel</td>
<td>0.56</td>
</tr>
<tr>
<td>sheep</td>
<td>0.55</td>
</tr>
<tr>
<td>pigeon</td>
<td>0.54</td>
</tr>
<tr>
<td>deer</td>
<td>0.53</td>
</tr>
<tr>
<td>rat</td>
<td>0.53</td>
</tr>
<tr>
<td>bird</td>
<td>0.53</td>
</tr>
</tbody>
</table>

#### good (window size= 2)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td>0.68</td>
</tr>
<tr>
<td>excellent</td>
<td>0.66</td>
</tr>
<tr>
<td>superb</td>
<td>0.48</td>
</tr>
<tr>
<td>poor</td>
<td>0.45</td>
</tr>
<tr>
<td>improved</td>
<td>0.43</td>
</tr>
<tr>
<td>improve</td>
<td>0.43</td>
</tr>
<tr>
<td>perfect</td>
<td>0.42</td>
</tr>
<tr>
<td>clever</td>
<td>0.42</td>
</tr>
<tr>
<td>terrific</td>
<td>0.42</td>
</tr>
<tr>
<td>lucky</td>
<td>0.41</td>
</tr>
<tr>
<td>smashing</td>
<td>0.41</td>
</tr>
<tr>
<td>improving</td>
<td>0.41</td>
</tr>
<tr>
<td>wonderful</td>
<td>0.41</td>
</tr>
</tbody>
</table>
BLESS
Baroni and Lenci (2011)

- BLESS is formed by 26,554 tuples expressing a relation between a (target) concept and a relatum (concept)
  - 200 basic-level nominal concrete concepts, 8 relation types, each instantiated by multiple relata (nouns, verbs or adjectives)
  - relata extracted from various resources (WordNet, ConceptNet, Wikipedia, corpora, etc.)

<table>
<thead>
<tr>
<th>target concept</th>
<th>relation</th>
<th>relata</th>
</tr>
</thead>
<tbody>
<tr>
<td>rabbit</td>
<td>HYPER</td>
<td>animal, chordate, mammal, ...</td>
</tr>
<tr>
<td>guitar</td>
<td>COORD</td>
<td>violin, trumpet, piano, ...</td>
</tr>
<tr>
<td>beaver</td>
<td>MERO</td>
<td>fur, head, tooth, ...</td>
</tr>
<tr>
<td>sword</td>
<td>ATTRI</td>
<td>dangerous, long, heavy, ...</td>
</tr>
<tr>
<td>butterfly</td>
<td>EVENT</td>
<td>fly, catch, flutter, ...</td>
</tr>
<tr>
<td>villa</td>
<td>RAN.N</td>
<td>disease, assistance, game, ...</td>
</tr>
<tr>
<td>donkey</td>
<td>RAN.V</td>
<td>coincide, express, vent, ...</td>
</tr>
<tr>
<td>hat</td>
<td>RAN.J</td>
<td>quarterly, massive, obvious, ...</td>
</tr>
</tbody>
</table>
DSMs and Semantic Relations

ContentWindow20

COORD HYPER MERO ATTRI EVENT RAN.N RAN.J RAN.V

-2 -1 0 1 2
Pattern-based approaches (Hearst 1992)

- Select as linguistic contexts, **lexico-syntactic patterns** that express a given semantic relation between lexical items.
  - **Hypernymy**: \( x \) is a kind of \( y \)
    - \( y \) such as \( x \)
  - **Antonymy**: \( x \) but not \( y \)
    - \( x \) or \( y \)

- The pair \( \langle a, b \rangle \) is an instance of relation \( R \), if \( a \) and \( b \) are frequently linked by the patterns expressing \( R \).

- select instance pairs of a given semantic relation $R$
  - $Hypernymy = \{ \langle \text{dog, animal} \rangle, \langle \text{tulip, flower} \rangle, \langle \text{cypress, tree} \rangle \ldots \}$
  - $\langle a, b \rangle$ is an instance of relation $R$, if $\langle a, b \rangle$ is analogically similar to the instances of $R$
    - $\text{dog:animal} = \text{car:vehicle}$
Measure **relational similarity** between word pairs in a pair-pattern co-occurrence matrix

\[
\begin{pmatrix}
p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \\
4 & 0 & 3 & 2 & 1 & \vdots & 2 & 3 \\
0 & 1 & 2 & 0 & 3 & \vdots & 1 & 0 \\
1 & 2 & 3 & 1 & 0 & \vdots & 2 & 0 \\
\end{pmatrix}
\]
Pairs of words sharing a particular relation are related by the same constant offset between their neural embeddings.

- Distributional vectors built with a Recursive Neural Network.

\[
\text{animal} - \text{dog} + \text{tulip} = \text{flower}
\]
Most of existing approaches are partially supervised
  - pattern selections, “seed” instances of relations, etc.

Modeling semantic relations requires us to explain the **entailments** they license
  - $X$ is a dog $\Rightarrow$ $X$ is an animal
  - $X$ is a dog $\not\Rightarrow$ $X$ is a cat
Most of existing approaches are partially supervised
- pattern selections, “seed” instances of relations, etc.

Modeling semantic relations requires us to explain the entailments they license
- $X$ is a dog $\Rightarrow$ $X$ is an animal
- $X$ is a dog $\not\Rightarrow$ $X$ is a cat
Hypernymy in Distributional Semantics

- Hypernymy is an **asymmetric** relation
  - $X$ is a dog $\Rightarrow$ $X$ is an animal
  - $X$ is an animal $\not\Rightarrow$ $X$ is a dog

- Hypernyms are **semantically broader** terms than their hyponyms
  - extensionally broader
    - *animal* refers to a broader set of entities than *dog*
  - intensionally broader
    - *animal* has more general properties than *dog* (e.g. bark)
    - superordinates are *less informative* than basic level concepts (Murphy 2002)
Hypernymy in Distributional Semantics

- Hypernymy is an asymmetric relation
  - $X$ is a dog $\Rightarrow$ $X$ is an animal
  - $X$ is an animal $\not\Rightarrow$ $X$ is a dog

- Hypernyms are semantically broader terms than their hyponyms
  - extensionally broader
    - *animal* refers to a broader set of entities than *dog*
  - intensionally broader
    - *animal* has more general properties than *dog* (e.g. bark)
    - superordinates are *less informative* than basic level concepts (Murphy 2002)
Since the class (extension) denoted by a hyponym is included in the extension denoted by the hypernym, hyponyms are expected to occur in a subset of the contexts of their hypernyms.

**Distributional Inclusion Hypothesis (DIH) (Kotlerman et al. 2010)**

If $u$ is a semantically narrower term than $v$, then a significant number of salient distributional features of $u$ is included in the feature vector of $v$ as well.
Directional Similarity Measures Based on the DIH

**WeedsPrec** (Weeds & Weir, 2003; Weeds et al., 2004)

WeedsPrec(u, v) = \[
\frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)}
\]  \hspace{1cm} (1)

**ClarkeDE** (Clarke 2009)

ClarkeDE(u, v) = \[
\frac{\sum_{f \in F_u \cap F_v} \min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}
\]  \hspace{1cm} (2)
Distributional Memory (Baroni and Lenci 2010)

- Distributional features are syntactically typed collocates: 
  \textit{subj\_intr\_sing}, \textit{obj\_read}, \textit{subj\_tr\_read}, etc.

- The context weighting function is \textbf{Positive Local Mutual Information} (LMI)

- The Distributional Memory corpus
  - \textbf{2.830 billion} tokens resulting from concatenating
    - \textit{ukWac}, about 1.915 billion tokens of Web-derived texts
    - \textit{English Wikipedia}, a mid-2009 dump of about 820 million tokens
    - \textit{British National Corpus}, about 95 million tokens

- the corpus was tokenized, POS-tagged and lemmatized with the TreeTagger, and dependency-parsed with the MaltParser
Directional Similarity Measures on BLESS

WeedsPrec

Alessandro Lenci
Directional Similarity Measures on BLEXS

ClarkeDE

coord  hyper  mero  random-n

-1.5  -1.0  -0.5  0.0  0.5  1.0  1.5

Alessandro Lenci
Directional Similarity Measures on BLESS

Relatum Frequency
Evaluation with Average Precision (AP)

- AP combines precision, relevance ranking and overall recall
- For each similarity measure, AP is computed with respect to the 4 BLESS relations
  - the best possible score (AP = 1) for a given relation (e.g., HYPER) corresponds to the ideal case in which all the relata belonging to that relation have higher similarity scores than the relata belonging to the other relations
- For each relation $R$, AP is computed for each of the 200 BLESS target concepts

$$AP(u, R) = \sum_{r=1}^{\left| R \right|} \left( \frac{P(r) \times rel(r))}{\left| R \right|} \right)$$

(3)

$$rel(r) = \begin{cases} 1 & \text{if the word at rank } r \text{ has a relation } R \text{ with } u \\ 0 & \text{otherwise} \end{cases}$$

(4)
Evaluation with Average Precision (AP)

<table>
<thead>
<tr>
<th>measure</th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.79</td>
<td>0.23</td>
<td>0.21</td>
<td>0.30</td>
</tr>
<tr>
<td>WeedsPrec</td>
<td>0.45</td>
<td>0.40</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td>ClarkeDE</td>
<td>0.45</td>
<td>0.39</td>
<td>0.28</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Mean AP values for each semantic relation reported by the different similarity scores
New directional similarity measure
Lenci and Benotto (2012)

\textit{invCL} (inverse ClarkeDE)

\begin{equation}
\text{invCL}(u, v) = \sqrt{\text{ClarkeDE}(u, v) * (1 - \text{ClarkeDE}(v, u))}
\end{equation}

- A broader term should also be found in contexts in which the narrow term is not used
- If \( v \) is a semantically broader term of \( u \), then the features of \( u \) are included in the features of \( v \), but the features of \( v \) are also not included in the features of \( u \), so that:
  1. a significant number of the \( u \)-contexts are also \( v \)-contexts
  2. a significant number of \( v \)-contexts are not \( u \)-contexts.
New directional similarity measure
Lenci and Benotto (2012)

*invCL* (inverse ClarkeDE)

\[
invCL(u, v) = \sqrt{\text{ClarkeDE}(u, v) \ast (1 - \text{ClarkeDE}(v, u))}
\]  

- A broader term should also be found in contexts in which the narrow term is *not* used.
- If \( v \) is a semantically broader term of \( u \), then the features of \( u \) are included in the features of \( v \), but the features of \( v \) are also *not* included in the features of \( u \), so that:
  1. a significant number of the \( u \)-contexts are also \( v \)-contexts
  2. a significant number of \( v \)-contexts are not \( u \)-contexts.
Directional similarity measures
### Evaluation with Average Precision (AP)

<table>
<thead>
<tr>
<th>measure</th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.79</td>
<td>0.23</td>
<td>0.21</td>
<td>0.30</td>
</tr>
<tr>
<td>WeedsPrec</td>
<td>0.45</td>
<td>0.40</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td>ClarkeDE</td>
<td>0.45</td>
<td>0.39</td>
<td>0.28</td>
<td>0.33</td>
</tr>
<tr>
<td>invCL</td>
<td>0.38</td>
<td>0.40</td>
<td>0.31</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Mean AP values for each semantic relation reported by the different similarity scores.
The intension (concept) expressed by a hypernym includes more general properties than the intension of its hyponyms

- *animal*: move, eat, is alive, etc.
- *dog*: bark, has fur, has four legs, etc.

**Distributional Informativeness Hypothesis (DIH) (Santus et al. 2014)**

The most typical linguistic contexts of a hypernym are less informative than the most typical linguistic contexts of its hyponyms.
For every word $w$ we identify the $N$ most associated contexts $c$

For each selected context $c$ we calculate its entropy:

$$H(c) = \sum_{i=1}^{n} p(c, f_i) \log_2 p(c, f_i)$$

For each $w$ we calculate a generality index $E_w$ as the median among the entropies of its $N$ contexts

We compare the semantic generality between two words $w_1$ and $w_2$:

$$SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}}$$

If $SLQS(w_1, w_2) > 0$, then $w_1$ is semantically less general than $w_2$
An Intensional Approach to Hypernymy in DSMs
Santus et al. (2014)

- **Experiment 1** - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair \((w_1, w_2)\), the hypernym is correctly identified iff \(SLSQ(w_1, w_2) > 0\)

<table>
<thead>
<tr>
<th></th>
<th>SLQS</th>
<th>baseline (freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>87.00%</td>
<td>66.09%</td>
</tr>
</tbody>
</table>

- **Experiment 2** - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

<table>
<thead>
<tr>
<th></th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>baseline</strong></td>
<td>0.51</td>
<td>0.40</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>SLQS*Cosine</strong></td>
<td>0.27</td>
<td>0.59</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>
An Intensional Approach to Hypernymy in DSMs
Santus et al. (2014)

- **Experiment 1** - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair \((w_1, w_2)\), the hypernym is correctly identified iff \(SLSQ(w_1, w_2) > 0\)

<table>
<thead>
<tr>
<th></th>
<th>SLQS</th>
<th>baseline (freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>87.00%</td>
<td>66.09%</td>
</tr>
</tbody>
</table>

- **Experiment 2** - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

<table>
<thead>
<tr>
<th></th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.51</td>
<td>0.40</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td>SLQS*Cosine</td>
<td>0.27</td>
<td>0.59</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>
An Intensional Approach to Hypernymy in DSMs
Santus et al. (2014)

- **Experiment 1** - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS
  - given a hyponym - hypernym pair \((w_1, w_2)\), the hypernym is correctly identified iff \(\text{SLSQ}(w_1, w_2) > 0\)

<table>
<thead>
<tr>
<th></th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.51</td>
<td>0.40</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td>SLQS*Cosine</td>
<td>0.27</td>
<td>0.59</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>

- **Experiment 2** - discriminating hypernyms from other semantic relations
  - SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

<table>
<thead>
<tr>
<th></th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.51</td>
<td>0.40</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td>SLQS*Cosine</td>
<td>0.27</td>
<td>0.59</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>
An Intensional Approach to Hypernymy in DSMs
Santus et al. (2014)

**Experiment 1** - identifying the hypernym in the 1277 hypernymy-related pairs in BLESS

- given a hyponym - hypernym pair \((w_1, w_2)\), the hypernym is correctly identified iff \(SLSQ(w_1, w_2) > 0\)

<table>
<thead>
<tr>
<th></th>
<th>SLQS</th>
<th>baseline (freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>87.00%</td>
<td>66.09%</td>
</tr>
</tbody>
</table>

**Experiment 2** - discriminating hypernyms from other semantic relations

- SLQS is combined here with the cosine (hypernyms are both very similar but differ for semantic generality)

<table>
<thead>
<tr>
<th></th>
<th>COORD</th>
<th>HYPER</th>
<th>MERO</th>
<th>RANDOM-N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>baseline</strong></td>
<td>0.51</td>
<td>0.40</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>SLQS*Cosine</strong></td>
<td>0.27</td>
<td>0.59</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>
the reality of contextual representations has been argued so far with only a bare mention of the most likely source of counter-examples, namely, *antonymous adjectives*. The problem is easily stated. Antonyms have contrasting meanings - not just zero semantic similarity, but negative similarity, if that is possible. Yet they seem to be freely substitutable for one another: If the referent of a head noun has a particular attribute, the noun could take either polar value of that attribute.


- Cosine similarity is not able to discriminate between synonyms and antonyms

<table>
<thead>
<tr>
<th></th>
<th>good</th>
<th>bad</th>
<th>excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td></td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>bad</td>
<td>0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>excellent</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The reality of contextual representations has been argued so far with only a bare mention of the most likely source of counter-examples, namely, antonymous adjectives. The problem is easily stated. Antonyms have contrasting meanings - not just zero semantic similarity, but negative similarity, if that is possible. Yet they seem to be freely substitutable for one another: If the referent of a head noun has a particular attribute, the noun could take either polar value of that attribute.


- Cosine similarity is not able to discriminate between synonyms and antonyms

<table>
<thead>
<tr>
<th></th>
<th>good</th>
<th>bad</th>
<th>excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.68</td>
<td>0.66</td>
</tr>
</tbody>
</table>
The Antonymy Conundrum

Antonyms are strongly associated in the mental lexicon (Deese 1964, 1965)

The Co-Occurrence Hypothesis (Miller and Charles 1989, Fellbaum 1995)
Semantically opposed lexemes tend to co-occur in the same sentences

Many empirical validations (Fellbaum 1995, Jones et al. 2012):

- All creatures great and small
- Rich and poor alike
- A matter of life or death
- Will the danger increase or decrease!
The Antonymy Conundrum

Antonyms are strongly associated in the mental lexicon (Deese 1964, 1965)

The Co-Occurrence Hypothesis (Miller and Charles 1989, Fellbaum 1995)
Semantically opposed lexemes tend to co-occur in the same sentences

Many empirical validations (Fellbaum 1995, Jones et al. 2012):

- *All creatures* great and small
- *Rich and poor* alike
- *A matter of* life or death
- *Will the danger* increase or decrease!
Mohammad et al. (2012) use an analogy-based approach combined with the Co-Occurrence Hypothesis.

Kim et al. (2013) apply neural embeddings to reconstruct intermediate values (e.g., angry) in adjective scales, given the antonyms (e.g., furious – happy).
Antonyms in Distributional Semantics

- The Co-Occurrence Hypothesis is not enough to identify antonyms
  - If two words are antonyms, they tend to co-occur, but...
    - ...many words pairs that tend to co-occur are not antonyms

- Modeling antonyms requires us to explain the entailments they license

  **complementary**
  - \(X\) is a alive \(\Rightarrow\) \(X\) is not dead
  - \(X\) is not alive \(\Rightarrow\) \(X\) is dead

  **contrary**
  - \(X\) is furious \(\Rightarrow\) \(X\) is not happy
  - \(X\) is not furious \(\not\Rightarrow\) \(X\) is happy
The Co-Occurrence Hypothesis is not enough to identify antonyms

If two words are antonyms, they tend to co-occur, but...

...many words pairs that tend to co-occur are not antonyms

Modeling antonyms requires us to explain the entailments they license

- **complementary**
  - $X$ is a alive $\Rightarrow$ $X$ is not dead
  - $X$ is not alive $\Rightarrow$ $X$ is dead

- **contrary**
  - $X$ is furious $\Rightarrow$ $X$ is not happy
  - $X$ is not furious $\nRightarrow$ $X$ is happy
Antonyms in Distributional Semantics

- The Co-Occurrence Hypothesis is not enough to identify antonyms
  - If two words are antonyms, they tend to co-occur, but...
    ...many words pairs that tend to co-occur are not antonyms

- Modeling antonyms requires us to explain the **entailments** they license
  
  **complementary**
  
  \[ \text{X is a alive} \Rightarrow \text{X is not dead} \]
  
  \[ \text{X is not alive} \Rightarrow \text{X is dead} \]

  **contrary**
  
  \[ \text{X is furious} \Rightarrow \text{X is not happy} \]
  
  \[ \text{X is not furious} \not\Rightarrow \text{X is happy} \]
Adjectival Antonyms in Semantic Spaces
Benotto and Lenci (2014)

Distributional Negation Hypothesis

If two adjectives are antonyms, each adjective is distributionally similar to the negation of the other

- The adjective *alive* is expected to share many contexts with NOT-*dead*
- Adjectives and their negation are represented with distributional vectors
  - $w^+$ vector derived from all the positive occurrences of $w$ in the training corpus
    - e.g. *The wood seemed alive, yet silent except for a wren*
  - $w^-$ vector derived from all the negative occurrences of $w$ in the training corpus
    - e.g. *The patient is not alive in the morning*
Adjectival Antonyms in Semantic Spaces
Benotto and Lenci (2014)

Hypothesis if \( w_i \) and \( w_j \) are antonyms,
\[
(\text{cosine}(w_i^+, w_j^-) \lor (\text{cosine}(w_i^-, w_j^+))) > \text{cosine}(w_i^+, w_j^+)
\]

Test set 83 pairs of antonyms (frequency of each antonym and its negation > 50)

<table>
<thead>
<tr>
<th>automatic</th>
<th>conscious</th>
</tr>
</thead>
<tbody>
<tr>
<td>comfortable</td>
<td>uncomfortable</td>
</tr>
<tr>
<td>obligatory</td>
<td>optional</td>
</tr>
<tr>
<td>drunk</td>
<td>sober</td>
</tr>
<tr>
<td>horizontal</td>
<td>vertical</td>
</tr>
<tr>
<td>national</td>
<td>international</td>
</tr>
</tbody>
</table>

Evaluation Accuracy of 77.10%
Adjectival Antonyms in Semantic Spaces
Benotto and Lenci (2014)

Hypothesis if \( w_i \) and \( w_j \) are antonyms,
\[
(cosine(w_i^+, w_j^-) \lor (cosine(w_i^-, w_j^+))) > cosine(w_i^+, w_j^+)
\]

Test set 83 pairs of antonyms (frequency of each antonym and its negation > 50)

- automatic, conscious
- comfortable, uncomfortable
- obligatory, optional
- drunk, sober
- horizontal, vertical
- national, international

Evaluation Accuracy of 77.10%
**Adjectival Antonyms in Semantic Spaces**

Benotto and Lenci (2014)

**Hypothesis**  
If $w_i$ and $w_j$ are antonyms,  
$$(\cosine(w_i^+, w_j^-) \lor \cosine(w_i^-, w_j^+)) > \cosine(w_i^+, w_j^+)$$

**Test set**  
83 pairs of antonyms (frequency of each antonym and its negation > 50)

- automatic, conscious
- comfortable, uncomfortable
- obligatory, optional
- drunk, sober
- horizontal, vertical
- national, international

**Evaluation**  
Accuracy of 77.10%
Conclusions

Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics.
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model.
  - We have barely scratched the surface of key phenomena like compositionality, inference, semantic relations, etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories.
  - Context-sense shifts, gradience, usage-based aspects of meaning, etc.
Conclusions

Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics.
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model.
  - We have barely scratched the surface of key phenomena like compositionality, inference, semantic relations etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories.
  - Context-sense shifts, gradience, usage-based aspects of meaning, etc.
Will Distributional Semantics Ever Become Semantics?

- Distributional semantics was born as an empirical method to measure semantic similarity with corpus-based distributional statistics.
- The range of addressed phenomena has enormously increased, but distributional semantics has still to prove its plausibility as a general semantic model.
  - We have barely scratched the surface of key phenomena like compositionality, inference, semantic relations etc.
- Distributional semantics is able to address phenomena that are very problematic for other semantic theories.
  - Context-sense shifts, gradience, usage-based aspects of meaning, etc.
Will Distributional Semantics Ever Become Semantics?

Two major issues:
- To what extent semantic phenomena can be formulated in terms of vector similarity?
- Which aspects of semantic knowledge can be decoded from linguistic distributions?

New prospectives come from integrating distributional semantics with other models of meaning, to establish a division of semantic labour that capitalizes on complementary strengths.
Will Distributional Semantics Ever Become Semantics?

Two major issues:

- To what extent semantic phenomena can be formulated in terms of vector similarity?
- Which aspects of semantic knowledge can be decoded from linguistic distributions?

New prospectives come from integrating distributional semantics with other models of meaning, to establish a division of semantic labour that capitalizes on complementary strengths.
Conclusions

Computational Models of Language Meaning in Context
Dagstuhl Seminar, November 10 – 15, 2013

Organizers    Hans Kamp, Alessandro Lenci, James Pustejovsky
Aims          Probing the limits of distributional semantics and fostering new synergies with other semantic frameworks
Collaborators

Marco Baroni, Giulia Benotto, Gianluca Lebani, Qin Lu, Magnus Sahlgren, Enrico Santus, Sabine Schulte im Walde
Thank You!!!
Tänan!!!
Grazie!!!


J. Deese (1965), *The Structure of Associations in Language and Thought*, Johns Hopkins Press, Baltimore, MD


Z. S. Harris (1951), *Methods in Structural Linguistics*, University of Chicago Press, Chicago, IL


References


