From Word Co-Occurrences to Properties of Concepts: Using corpora to simulate the human experience

Marco Baroni

Center for Mind/Brain Sciences (University of Trento)

Bressanone/Brixen
March 28, 2008
Collaborators

- Brian Murphy, Eduard Barbu, Massimo Poesio (CIMeC)
- Alessandro Lenci (University of Pisa and ILC/CNR)
- Marco Baroni, Eduard Barbu, Brian Murphy and Massimo Poesio, in preparation. StruDEL: A distributional semantic model based on properties and types
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
Corpora and the human experience

- Computational modelers in cognitive science (e.g., Rogers and McClelland 2004) typically work with hand-crafted input
- Corpora are “real”, natural input, akin to what humans hear/read, with same problems of noise and skewed input distribution (Zipf’s law) humans must face
The contextual view meaning

- Acquisition/representation of meaning/conceptual knowledge is core issue in cognitive science
- Corpus-based simulations can help!
The contextual view meaning

- Acquisition/representation of meaning/conceptual knowledge is core issue in cognitive science
- Corpus-based simulations can help!
- “You should tell a word by the company it keeps” (Firth, 1957)
- “[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts” (Cruse, 1986)
Outline

Introduction

**Distributional semantics**

From distributional semantics to conceptual knowledge

**StruDEL**

Testing StruDEL
  - Property generation
  - Categorization
  - Other tasks

The Human Experience
Distributional semantics
Word space models (WSMs)

- Meaning of word/concept defined by *set of contexts* in which word occurs in corpus
- Similarity of words represented as *geometric distance* among *context vectors*
Distributional semantics
Co-occurrence extraction

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark
park
owner
leash
The dog barked in the park. The owner of the dog put him on the leash since he barked.
Distributional semantics
Co-occurrence extraction

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark | +
park | +
owner
leash
The dog barked in the park. The owner of the dog put him on the leash since he barked.
The dog barked in the park. The owner of the **dog** put him on the **leash** since he barked.
The dog barked in the park. The owner of the **dog** put him on the leash since he **barked**.

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bark</td>
<td>++</td>
</tr>
<tr>
<td>park</td>
<td>+</td>
</tr>
<tr>
<td>owner</td>
<td>+</td>
</tr>
<tr>
<td>leash</td>
<td>+</td>
</tr>
</tbody>
</table>
### Distributional semantics

Meaning as co-occurrence

<table>
<thead>
<tr>
<th></th>
<th>leash</th>
<th>walk</th>
<th>run</th>
<th>owner</th>
<th>pet</th>
<th>bark</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>cat</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</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>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Distributional semantics
Similarity in word space

![Diagram showing word space with 'owner' on the horizontal axis and 'pet' on the vertical axis. Points for 'dog (5,3)', 'cat (2,3)', and 'car (3,0)' are marked.]
Distributional semantics

Similarity in word space
Distributional semantics

Which context?

- Documents/large textual spans
- All words in a narrow window
- Lemmatized content words in a narrow window
- Content words in specific syntactic constructions or specific surface patterns
- Context needs not be linguistic! Vectors could include, e.g., co-occurrence counts with sensory stimuli
Distributional semantics
Success in cognitive simulations

- synonym identification (Landauer and Dumais 1997)
- text coherence (Landauer and Dumais 1997)
- categorization (Burgess and Lund 1997)
- semantic priming (Lowe 2000, McDonald and Brew 2002, Vigliocco et al. 2004)
- substitution errors (Vigliocco et al. 2004)
- child lexicon acquisition (Li et al. 2004, Baroni et al. 2007)
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
The flat model of semantic similarity

“Semantic similarity” is multi-faceted notion but a single WSM provides only one way to rank a set of words (according to distance measure of choice)

Nearest neighbours of *motorcycle* in a standard WSM:
- motor → component
- car → co-hyponym
- diesel → component?
- to race → proper function
- van → co-hyponym
- bmw → hyponym
- to park → proper function
- vehicle → hypernym
- engine → component
- to steal → frame?
“Although jugs might be related to both vinegar and bottles, these relations are extremely different, and an overall similarity score does not represent these differences.”

In order to distinguish how jugs are related to vinegar from how they are related to bottles, one needs to know what are the properties of these concepts:

“[S]ince one’s concept of a jug, say, would include detailed information about its origins, parts, materials, functions and so on, the concept is more than sufficient to distinguish the meaning of jugs from that of vinegar and, for that matter, bottles.”
Semantic relations in cognitive and applied tasks

- Property generation: humans can easily produce coherent lists of typical properties of concepts (*norms* of McRae et al., Vinson/Vigliocco and others)

- Humans are able to distinguish different *types* of relations between properties and concepts, e.g., between formal and functional properties
  - A *dish* looks like a *CD* but its function is more similar to that of a *bottle*
Semantic relations in cognitive and applied tasks

► Different relations must be extracted and identified for modeling semantic interpretation, e.g.:
  ► Telic quale relation in type coercion: finish the book (*to read*) vs. the ice-cream (*to eat*) (Pustejovsky 1995)
  ► Salient properties in compound interpretation: a *zebra cup* is a cup *with stripes*
  ► Parts in co-reference bridging: The *building* faced a dark alley. A *window* opened

► Specific *types* of properties (e.g., visual vs. functional) play crucial role in neural organization of concepts and semantic deficits (Martin 2007, Vigliocco et al. 2004)

► Semantic relations needed in practical applications, in particular development of lexical resources
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
StruDEL
Structured Dimensions Extraction and Labeling

- Tries to build contextual vectors that represent *typed properties of concepts*
Tries to build contextual vectors that represent *typed properties of concepts*

**Concept:** *motorcycle*

**(Target) representation:**
- *for* traveling
- *for* running
- *is-a* vehicle
- *has* motor
- *has* two wheels
- ...
Strategies for typed property extraction

- Automated identification of plausible concept-property connectors:
  - lice in a large number of dogs → YES
  - lice and leeches → NO

- (Weighted) number of distinct connectors between concept and property is better indicator of true semantic relation
- year of the tiger is very frequent, but following are not attested: year of some tigers, the tigers have years, etc.
- Vice versa, no tail/tiger connector is very frequent, but there are many of them: tail of the tiger, tail of some tigers, the tigers have tails, etc.

- The type of a relation can be extracted by generalizing over the connectors: of, with, of some, have, point, together, to a part/whole relation
Strategies for typed property extraction

- Automated identification of plausible concept-property connectors:
  - lice in a large number of dogs → YES
  - lice and leeches → NO
- (Weighted) number of distinct connectors between concept and property is better indicator of true semantic relation than absolute co-occurrence frequency
  - year of the tiger is very frequent, but following are not attested: year of some tigers, the tigers have years, etc.
  - Vice versa, no tail/tiger connector is very frequent, but there are many of them: tail of the tiger, tail of some tigers, the tigers have tails, etc.
Strategies for typed property extraction

- Automated identification of plausible concept-property connectors:
  - lice in a large number of dogs → YES
  - lice and leeches → NO

- (Weighted) number of distinct connectors between concept and property is better indicator of true semantic relation than absolute co-occurrence frequency
  - year of the tiger is very frequent, but following are not attested: year of some tigers, the tigers have years, etc.
  - Vice versa, no tail/tiger connector is very frequent, but there are many of them: tail of the tiger, tail of some tigers, the tigers have tails, etc.

- The type of a relation can be extracted by generalizing over the connectors: of, with, of some, have point, together, to a part/whole relation
Concept-property-type tuples extracted from ukWaC, a corpus of random Web pages including 2.25 billion tokens

Property lists extracted for 1,234 (concrete) concepts

Compared to various state-of-the-art WSMs, including SVD-based model and model using dependency parses (DV)
book
The StruDEL description

<table>
<thead>
<tr>
<th>property</th>
<th>type</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>to read</td>
<td>verb obj</td>
<td>3941.3</td>
</tr>
<tr>
<td>author</td>
<td>c from p</td>
<td>3772.8</td>
</tr>
<tr>
<td>to write</td>
<td>verb obj</td>
<td>2399.5</td>
</tr>
<tr>
<td>reader</td>
<td>c for p</td>
<td>2298.5</td>
</tr>
<tr>
<td>chapter</td>
<td>p in c</td>
<td>2259.8</td>
</tr>
<tr>
<td>library</td>
<td>c in p</td>
<td>2222.4</td>
</tr>
<tr>
<td>to publish</td>
<td>verb obj</td>
<td>1907.7</td>
</tr>
<tr>
<td>reading</td>
<td>p from c</td>
<td>1296.8</td>
</tr>
<tr>
<td>publisher</td>
<td>c from p</td>
<td>1258.0</td>
</tr>
<tr>
<td>review</td>
<td>p about c</td>
<td>1156.4</td>
</tr>
</tbody>
</table>
The StruDEL description

<table>
<thead>
<tr>
<th>property</th>
<th>type</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>to ride</td>
<td>verb obj</td>
<td>345.8</td>
</tr>
<tr>
<td>rider</td>
<td>p on c</td>
<td>199.8</td>
</tr>
<tr>
<td>vehicle</td>
<td>p as c</td>
<td>103.7</td>
</tr>
<tr>
<td>motorbike</td>
<td>p for c</td>
<td>100.2</td>
</tr>
<tr>
<td>street</td>
<td>c on p</td>
<td>71.3</td>
</tr>
<tr>
<td>to park</td>
<td>verb obj</td>
<td>69.3</td>
</tr>
<tr>
<td>scooter</td>
<td>p over c</td>
<td>51.6</td>
</tr>
<tr>
<td>car</td>
<td>p as c</td>
<td>45.7</td>
</tr>
<tr>
<td>to insure</td>
<td>p for c</td>
<td>39.8</td>
</tr>
<tr>
<td>bike</td>
<td>p out c</td>
<td>37.7</td>
</tr>
</tbody>
</table>
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
- 725 participants rating 541 concepts, 30 subjects per concept
- Subjects produce list of properties that describe concept
- Manual normalization: *loud, noise, noisy* mapped to *is loud*
On average, NORMS and StruDEL share 2.4 of the 10 most salient properties of a concept. No other distributional model we tested shares more than 1.5/10 properties with NORMS.
Property analysis

- On average, NORMS and StruDEL share 2.4 of the 10 most salient properties of a concept.
- No other distributional model we tested shares more than 1.5/10 properties with NORMS.
- Systematic analysis of different property types privileged by NORMS vs. StruDEL (Baroni and Lenci 2008).
- E.g., for car:
  - Shared: engine, gasoline, transportation
  - NORMS only: wheels, 4 wheels, doors, steering wheel, expensive, for passengers, vehicle
  - StruDEL only: it is driven, driver, it is parked, road, garage, race, parking
Property types
Wu and Barsalou (Submitted), McRae et al. (2005), simplified

C  Category: dog-animal, airplane-vehicle
P  Parts: dog-tail, airplane-wing
Q  Qualities: dog-brown, airplane-fast
A  Typical Activities and behaviours: dog-barks, airplane-flies
F  Function: dog-pet, dog-hunting, airplane-transportation
E  Related Entities: dog-cat, airplane-pilot
L  Location: dog-kennel, airplane-sky
Property types in NORMS and StruDEL

<table>
<thead>
<tr>
<th>Property Type</th>
<th>Semantic Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>StruDEL</strong></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>Q</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>L</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property Type</th>
<th>Semantic Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>NORMS</strong></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>Q</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>L</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property Type</th>
<th>Semantic Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Semantic Space</strong></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>Q</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>L</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property Type</th>
<th>Semantic Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>StruDEL</strong></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>Q</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>L</td>
</tr>
</tbody>
</table>
Properties by categories

NORMS
- vehicle
- tool
- veggie
- fruit
- grAnim
- bird

StruDEL
- vehicle
- tool
- veggie
- fruit
- grAnim
- bird

C P Q A F E L
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

**Testing StruDEL**

- Property generation
- Categorization
- Other tasks

The Human Experience
Data-set

44 concrete concepts

- 24 natural concepts
  - 15 animals: 7 birds, 8 ground animals
  - 9 vegetables: 4 fruits, 5 greens
- 20 artifacts
  - 13 tools
  - 7 vehicles
Hierarchical categorization

- 6-way: birds, ground animals, fruits, greens, tools, vehicles
- 3-way: animals, vegetables, man-made
- 2-way: natural, man-made
Categorization as clustering in semantic space
Categorization as clustering in semantic space
## Results

Percentage *purity* of clusters

<table>
<thead>
<tr>
<th>space</th>
<th>6 categories</th>
<th>3 categories</th>
<th>2 categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMS</td>
<td>91</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>StruDEL</td>
<td>79</td>
<td>91</td>
<td>98</td>
</tr>
<tr>
<td>DV</td>
<td>73</td>
<td>89</td>
<td>95</td>
</tr>
<tr>
<td>SVD</td>
<td>79</td>
<td>75</td>
<td>59</td>
</tr>
</tbody>
</table>
“Animals”
Typical properties in the 3-way solution

NORMS
- animal
- legs
- beak
- eggs
- bird

StruDEL
- it breeds
- is seen
- is shot
- is rescued
- dies
“Vegetables”
Typical properties in the 3-way solution

<table>
<thead>
<tr>
<th>NORMS</th>
<th>StruDEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>➤ vegetable</td>
<td>➤ is sliced</td>
</tr>
<tr>
<td>➤ sweet</td>
<td>➤ is minced</td>
</tr>
<tr>
<td>➤ on trees</td>
<td>➤ is eaten</td>
</tr>
<tr>
<td>➤ fruit</td>
<td>➤ it grows</td>
</tr>
<tr>
<td>➤ edible</td>
<td>➤ slice</td>
</tr>
</tbody>
</table>
“Man-made”
Typical properties in the 3-way solution

NORMS
- metal
- plastic
- handle
- for transportation
- wood

StruDEL
- is used
- in hands
- powered
- has use
- makes things
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
Other tasks
StruDEL is consistently best performer

- Predicting free association (*cat* . . . *dog*)
- Modeling prototypicality ratings (a *sparrow* is a more typical bird than a *penguin*)
- Generating specific properties (what is the typical *location* of hammers? what is their typical *function*?)
From word spaces to brain spaces

Work in progress

- Cat (2,3)
- Dog (5,3)
- Car (3,0)

Graph showing the projection of objects onto components.
Outline

Introduction

Distributional semantics

From distributional semantics to conceptual knowledge

StruDEL

Testing StruDEL
  Property generation
  Categorization
  Other tasks

The Human Experience
my brother
and I got dried <bananas> when war broke out
three pounds of <bananas> please
we could have a <banana> souffle
Those <bananas> are going a bit mm
well I like them when they go soft

who eats <bananas> ?
who do we have in
our zoo that eats <bananas> ?
monkeys eat <bananas>
shall we
give the monkey a <banana> ?
there’s your <banana>
someone went berserk with a <hammer>, that’s been known we had <hammer> drilled the blunt bit I was tapping it with a <hammer> wasn’t I

she had a <hammer> and she was banging on the wall the <hammer> is the most useful tool, Lara. Whenever the telly goes wrong you just hit it with your <hammer>
The Human Experience

- We can gain useful insights about human conceptual acquisition applying fancy learning techniques to whatever corpora we have available.
- But we suspect we will not make a major breakthrough until we learn from the same data humans learn from:
  - Corpora cue how adults transfer knowledge to other adults.
  - Knowledge in corpora is not organized incrementally.
  - Large corpora (still) lack multimodal information.
- Corpora in CHILDES are too small, sparse, often record speech in special occasions.
  - Lara, one of densest CHILDES corpora, contains transcripts of about 120 hours.
The Human Experience

► We can gain useful insights about human conceptual acquisition applying fancy learning techniques to whatever corpora we have available

► But we suspect we will not make a major breakthrough until we learn from the same data humans learn from:
  ▶ Corpora cue how adults transfer knowledge to other adults
  ▶ Knowledge in corpora is not organized incrementally
  ▶ Large corpora (still) lack multimodal information

► Corpora in CHILDES are too small, sparse, often record speech in special occasions
  ▶ Lara, one of densest CHILDES corpora, contains transcripts of about 120 hours

► The Human Experience: record full verbal and visual experience of multiple children uninterruptedly for first 3 years of life
Some references

Some more references

L. Wu & L. Barsalou (Submitted). *Grounding concepts in perceptual simulation: I. Evidence from property generation.*