Distributional semantics IV:
Is distributional semantics really “semantics”?

Marco Baroni

UPF Computational Semantics Course
Lecture strongly inspired by reading:

Everybody agrees that compositionality is a core aspect of natural language, and in particular natural language semantics.

<table>
<thead>
<tr>
<th>word</th>
<th>type</th>
<th>function</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>e</td>
<td>John</td>
<td>John</td>
</tr>
<tr>
<td>sings</td>
<td>e→t</td>
<td>λ x.sings(x)</td>
<td>{John, Mary}</td>
</tr>
</tbody>
</table>

John sings
sings(John) = TRUE

Compositionality allows us to create infinite meanings with finite means.
Compositionality and distributional semantics

- Vectors are supposed to represent the meaning of words
- What happens when words are composed?
- Should compositionality be represented in vector space?
- If not, is this a problem?
  - Long tradition of scholars unsympathetic to statistical approaches to language have argued that they are doomed to fail because they cannot capture compositionality (Chomsky, Pinker, Marcus…)
- Can vector spaces help in dealing with compositionality-related issues, without being the whole story?
Composition as vector combination

- Given two word vectors, the distributional “meaning” of the phrase combining them is given by a combined vector (e.g., the sum of their vectors)
Composition as vector combination

- It might work to some extent, since the vector of *car runs* will be somewhat similar to the vectors of *car*, *run* and related words.
- However, order/hierarchical structure is not taken into account.
  - *Run a car* and *The car runs* produce identical vectors.
- Leads to comparison of apples and oranges.
  - Is *The car runs* “more similar” to *The car costs*, to *The cat runs* or to *The car runs only with expensive gasoline*?
Making vector combination more sophisticated

- Jones and Mewhort (2007), following up on work by Smolensky (1990) and Plate (2003), propose sophisticated vector composition methods ("tensor algebra") where the resulting vector keeps track of the order of words in the original phrase.

- Very interesting developments, but for the moment:
  - It is not clear that they could ever possibly capture the richness of hierarchical relations.
  - Even if they will, we will still need theory of what combines with what and how: tensor algebra might provide the means to encode such combinations in a vector space.
Compositionality and word vectors

- A phrase might correspond to a single vector in special cases, e.g. concept combinations with idiomatic/specialized aspects of meaning.

- Most semanticists would however agree that when we compose words/concepts, we are stating a relation between concepts, rather than fusing them.
  - Although the combined concepts might influence each other, see e.g., Hampton’s work on combined concept prototypicality: a wooden spoon is prototypically larger than a spoon.

- Pursue an approach in which, when sentences are built, word vectors are checked and updated to enforce various composition constraints, but they are not fused.
Selectional preferences

- We saw that distributional semantics reaches reasonable performance in modeling selectional preferences (you eat topinambur, not sympathy).
- Given general semantic composition rules that combine words having very few types (e, t, etc.), vector-based distributional semantic techniques can be used to check the “commonsense” plausibility of the combination.
- The Montagovian composition component might tell us that both John eats topinambur and John eats sympathy are false.
- Distributional semantics will tell us that the latter is also highly unlikely.
Outline

Compositionality
  Co-composition
  Structured compositionality

More open issues in semantic distributional semantics
Co-composition
Pustejovsky 1995, and many others

- When words are composed, they tend to affect each other’s meanings
- *The horse runs* vs. *The water runs*
- “The horse horse-like runs”
- Erk and Padó (2008): the distributional vector of a predicate/argument in the context of an argument/predicate combines the distributional vector of the predicate/argument with the distributional vector of the prototypical predicate/argument of the argument/predicate
The *run* vector in the context of *horse* is a combination of the *run* vector and a prototype vector that represents the typical verbs *horse* is a subject of:

- Prototype vector constructed by (weighted) sum of vectors of words that are seen in the relevant relation to the target word (here, *horse+subj*<sup>-1</sup>), similarly to what I’ve shown for selectional preferences.
- Except that before combining, prototype vector must be normalized, or else it will be so long as to dominate the composition entirely.
Vector combination
before and after normalization
Vector combination

- Following Mitchell and Lapata (2008), Erk and Padó prefer component-wise vector multiplication to summing.

- Does not have a straightforward geometric interpretation, but it produces an “intersection” effect:

<table>
<thead>
<tr>
<th></th>
<th>in</th>
<th>with</th>
<th>in</th>
<th>with</th>
<th>under</th>
<th>obj</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>field</td>
<td>legs</td>
<td>race</td>
<td>gasoline</td>
<td>ground</td>
<td>forage</td>
</tr>
<tr>
<td>run</td>
<td>15.3</td>
<td>24.7</td>
<td>23.2</td>
<td>20.3</td>
<td>10.1</td>
<td>0.0</td>
</tr>
<tr>
<td>horse-subj$^{-1}$</td>
<td>8.9</td>
<td>15.2</td>
<td>20.1</td>
<td>0.0</td>
<td>1.2</td>
<td>24.5</td>
</tr>
<tr>
<td>summed</td>
<td>24.2</td>
<td>39.9</td>
<td>43.3</td>
<td>20.3</td>
<td>11.3</td>
<td>24.5</td>
</tr>
<tr>
<td>multiplied</td>
<td>136.17</td>
<td>375.44</td>
<td>466.32</td>
<td>0.00</td>
<td>12.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Predicting verb-in-context similarity judgments

- Mitchell & Lapata’s (2007, ML) data-set
- 49 subjects produced similarity ratings on 1-7 scale for intransitive subject-verb sentence pairs, one with context-affected verb, one with “landmark” reference verb:

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>landmark</th>
<th>judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>shoulder</td>
<td>slump</td>
<td>slouch</td>
<td>7</td>
</tr>
<tr>
<td>shoulder</td>
<td>slump</td>
<td>decline</td>
<td>2</td>
</tr>
<tr>
<td>value</td>
<td>slump</td>
<td>slouch</td>
<td>3</td>
</tr>
<tr>
<td>value</td>
<td>slump</td>
<td>decline</td>
<td>7</td>
</tr>
</tbody>
</table>

- Average inter-subject Spearman correlation ($\rho$): 40%!
Erk & Padó setup

- (I skip over many details and further analyses)
- Vectors computed on 100M word BNC corpus, MINIPAR-based dependency-links, MI-weighting
- Measure cosine similarity of vector that, according to various models, represent verb-in-context, to landmark verb vector (e.g., “slump-in-the-context-of-value” vector to \textit{decline} vector)
Erk & Padó setup
Verb-in-context vector construction

- **verb**: use verb out-of-context vector
- **prototype**: use prototype vector built from vectors of verbs that typically occur with noun as subject
- **combined**: multiply verb vector and noun-as-subject prototype verb vector
- **power-combined**: same, but values of dimensions of noun-as-subject prototype verb vector are raised to a power of 20
- **ML**: Mitchell and Lapata’s method: multiply noun and verb vector
Results
Correlation with human judgments

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>27%</td>
</tr>
<tr>
<td>power-combined</td>
<td>24%</td>
</tr>
<tr>
<td>prototype</td>
<td>16%</td>
</tr>
<tr>
<td>combined</td>
<td>13%</td>
</tr>
<tr>
<td>verb</td>
<td>8%</td>
</tr>
</tbody>
</table>

Similar results for a subset of the SEMEVAL07 lexical substitution task, where *power-combined* and *prototype* outperform *ML*. 
Co-compositionality: conclusions

- Results promising, but more empirical work needed (see also Erk & Padó 2009)
- Does co-compositionality always happen or are predicate and argument vectors only composed in special cases? If so, when?
- Can selectional restrictions be reduced to the same co-composition operations?
  - Composing the *eating* vector with the *sympathy-as-object* verb prototype gives an eating-in-the-context-of-sympathy vector that is nearly empty (recall intersective interpretation of combining by multiplication), or meaningless (summing all properties of eating and activities concerning sympathy)
Co-compositionality: conclusions

- Co-compositionality appears to be asymmetric: water-in-the-context of running is more similar to out-of-context water than running-in-the-context-of-water is to running in general – can we derive/account for this?
- Recursive application: in principle it should be straightforward (eat in “eat rotten topinambur” is eat-in-the-context-of-topinambur-in-the-context-of-rotten), but many devils will be in the details, as usual
Outline

Compositionality
  Co-composition
  Structured compositionality

More open issues in semantic distributional semantics
Structured compositionality
Almuhareb 2006, Poesio & Almuhareb 2008

- Originally, a co-reference problem:
  - A yellow car\textsubscript{i} was parked in the alley. . . . The expensive car\textsubscript{i} had been abandoned . . .
  - A yellow car\textsubscript{i} was parked in the alley. . . . *The red car\textsubscript{i} had been abandoned . . .
- Here, it is not sufficient to “compose” \textit{car} with its modifiers: we must also know which modifiers are compatible, which ones are not
Problem tackled by building a distributional semantic space based (also) on *attributes* and *values*:

<table>
<thead>
<tr>
<th>concept</th>
<th>attribute</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>colour</td>
<td>yellow</td>
</tr>
<tr>
<td>car</td>
<td>colour</td>
<td>red</td>
</tr>
<tr>
<td>car</td>
<td>cost</td>
<td>expensive</td>
</tr>
</tbody>
</table>

Semantic space should provide default values (we might assume cost of car is “expensive”, until context makes us update value of *cost* attribute)

and impose constraints on context modification (an attribute slot can be filled only once for each entity)
Attributes and values
Almuhareb 2006, Poesio & Almuhareb 2008

- Harvested from a very large corpus (the Web), originally with a purely pattern-based approach:
  attributes: the * of the NOUN is (“the color of the car is”)
  values: the * NOUN is (“the red car is”)
- Later, also using dependency information
- Semantic space that treats attributes and values as separate dimensions reaches very good performance in concept categorization tasks
- However, work on using patterns also to connect attributes and values (color and red) still preliminary, results not too promising (Almuhareb 2006)
- Adjectives are difficult for distributional models!
Outline

Compositionality
  Co-composition
  Structured compositionality

More open issues in semantic distributional semantics
Evidence suggests that we can group word/concept pairs by relational similarity types, e.g., finding that dog/animal, car/vehicle, etc. all instantiate a subordinate/superordinate (hyponym/hypernym) relation.

...but this does not tell us what inferences are supported by the fact that we know the concepts to be in this relation.

E.g., that if Christine is a car, then it is also a vehicle, but not vice-versa.

Can we also learn information about the inference properties of relations from corpora, or is this outside the scope of distributional semantics?

Some preliminary, not too encouraging results in Erk (2009).
Function words

- Formal semantics is mostly about function words (quantifiers, negation, wh-words...)
- Distributional semantics tends to ignore them as target elements, only uses them as linking elements, if at all
- Are they beyond the scope of distributional semantics, or are we not there, yet?
- Could distributional semantics be part of the language faculty in the broad sense, whereas quantifiers etc. are the lexical expressions of the language faculty in the narrow sense? (Hauser, Chomsky & Fitch, 2002)
Denotational semantics and distributional vectors

- If the semantics of natural language is denotational, where do our distributional vectors fit in?
- Two hypotheses:
  - They are an approximation to (mental representations of?) world knowledge
    - Both *Jack ate topinambur* and *Jack ate sympathy* map to the empty set, but the vectors tell us that the latter is also a highly unlikely event
  - Denotations are not denotations of words, but denotations of (possibly contextually modulated) vectors (Roberto Zamparelli)
    - The syntactic representation of *runs* in *The horse runs* is a vector denoting all things that run in a horse-like manner
The symbol grounding problem

- Distributional semantics might be the closest AI got to Searle’s famous Chinese Room (Glenberg & Robertson 2000)
  - A distributional semantic space might solve difficult categorization or analogy problems, but we would not say that it “understands” them
- However, there is no problem in thinking that at least some dimensions of our language- (and possibly vision-)based semantic spaces are grounded by links to the sensory-motor system
- Even hardcore “embodied cognition” theorists recognize a role for word associations (Barsalou et al. 2008)
That's all, folks!

and many thanks for being still here!