Statistical Semantics

Walter Kintsch

Building Meaning from Language
Tufts University, June 14-16, 2007
• Rational Theory of Mind: Infer mental structure from structure of linguistic corpus
• Statistical Semantics: Similarity between words, sentences, and texts ONLY
Statistical Methods for the Representation of Meaning

- Holograph Model: Jones & Mewhort (2007)
- Topics Model: Griffith, Steyvers & Tenenbaum (2007)
• For a large class of cases – though not for all – in which we employ the word “meaning” it can be defined thus: the meaning of a word is its use in language.  

  Wittgenstein (1953)

• A word is characterized by the company it keeps.  

  Firth (1957)

• Language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others.  

  Saussure (1915)
Latent Semantic Analysis:

- LSA as a map of meaning
- High-dimensional semantic space
- Text meaning as the sum of the words
- Limitations:
  - Learns only from words
  - Neglect of order and syntax
Handbook of
BEAGLE - Holograph Model
Jones & Mewhort (2007)

• A semantic memory model adapted after the episodic memory model of Murdock (1982)
• Distinguishes item and order information
• Item information for is the sum of all word vectors that have appeared with it in the same sentence
• Order information is the convolution of all word sequences that have appeared with it in the same sentence
Both item and order information are needed (Jones, Kintsch, & Mewhort, 2006):

- **semantic priming**
  - (deer-pony) >
- **associative priming**
  - (bee-honey):
- LSA & item alone fails
- item + order predicts

- **mediated priming**
  - (lion-stripes):
- HAL & order alone fails
- item + order predicts
**Topics Model**: Griffiths, Steyvers & Tenenbaum (2007)

**Independent Component Analysis**: Mangalath (2007)

- ICA assumes that each observation $x_i$ (word, document) is a mixture of independent semantic elements (components, topics):
  
  $x_i = a s_1 + b s_2 + \ldots + n s_n$

  or in matrix notation

  $x = As$

- $A$ is the mixing matrix; $s$ is the set of semantic elements; both are unknown and must be estimated at the same time
• ICA extracts 2,000 independent semantic components from the TASA corpus. Since the component probabilities are independent, the probability of a word \( W \) with the components \( \varepsilon_i \), \( i = 1, \ldots, 2,000 \) is

\[
p(W) = \sum_i p(W|\varepsilon_i)
\]

• The conditional probability of word \( W \) in the context of word \( C \) is

\[
p(W|C) = \frac{p(W \cap C)}{p(C)} = \frac{\sum_i \{\varepsilon_i \ p(W|\varepsilon_i) \ p(C|\varepsilon_i)\}}{\sum_i \ p(C|\varepsilon_i)}
\]
<table>
<thead>
<tr>
<th>20 nearest neighbors by LSA cosine</th>
<th>20 nearest neighbors by conditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>play</td>
<td>play</td>
</tr>
<tr>
<td>playing</td>
<td>game</td>
</tr>
<tr>
<td>played</td>
<td>playing</td>
</tr>
<tr>
<td>kickball</td>
<td>played</td>
</tr>
<tr>
<td>plays</td>
<td>fun</td>
</tr>
<tr>
<td>games</td>
<td>games</td>
</tr>
<tr>
<td>game</td>
<td>pat</td>
</tr>
<tr>
<td>volleyball</td>
<td>children</td>
</tr>
<tr>
<td>fun</td>
<td>ball</td>
</tr>
<tr>
<td>golf</td>
<td>role</td>
</tr>
<tr>
<td>costumes</td>
<td>plays</td>
</tr>
<tr>
<td>actor</td>
<td>important</td>
</tr>
<tr>
<td>rehearsals</td>
<td>music</td>
</tr>
<tr>
<td>actors</td>
<td>part</td>
</tr>
<tr>
<td>drama</td>
<td>run</td>
</tr>
<tr>
<td>comedy</td>
<td>friends</td>
</tr>
<tr>
<td>baseball</td>
<td>lot</td>
</tr>
<tr>
<td>tennis</td>
<td>stage</td>
</tr>
<tr>
<td>theater</td>
<td>toys</td>
</tr>
<tr>
<td>checkers</td>
<td>team</td>
</tr>
</tbody>
</table>
The architecture of memory:

• Holograph model:

• Multiple trace model:

• Is the difference between episodic and semantic memory structural or in the content encoded?
Connectionist Models:

- Anderson et al. (1977): categorical speech perception
  - Internal pattern of neuronal activity is consistent with multiple groups of neurons that cooperate with another to form a stable percept
Connectionist Models:

  – represents both item and order information

• McRae (2004): attractor networks
  – pattern of neuronal activity a location in a state space of which the dimensions are the activation levels of neurons
• **Advantages of statistical models:**
  *Computational theory - rational analysis*

• **Scale:**
  model is exposed to roughly the same amount of text as people encounter to match their semantic knowledge

• **Representativeness:**
  consequence of using an authentic linguistic corpus

• **Advantages of connectionist approach:**
  *Psychological process model*

  **Process** oriented (e.g., 0.5 sec for activation - as in priming data, such as Swinney, 1979)
Statistical models are plausible models of brain function:

- **LSA** - Hebbian learning
- **ICA and Holograph** - Gabor functions in visual system
The Representation of Meaning in Long Term Memory

- LSA, ICA, etc. generate decontextualized representations
- LSA, ICA etc. provide a map of meaning
- LSA, ICA etc. model what is stored in LTM
The Construction of Meaning in Working Memory

• Meaning is contextual
• Meaning is created when the map is used
• Meaning is generated in WM from the LTM representation + context
The Construction of Meaning in LSA

• **Predication Model** (Kintsch, 2001)
  • Context-relevant items from the semantic neighborhood of a word are selected
  • By a spreading activation process
  • Polysemy
  • *Noun-is-noun* type metaphors
  • Analogies
  • Asymmetric similarity judgments
The Construction of Meaning in ICA

- Distribution of semantic elements is modified: conditional probabilities
<table>
<thead>
<tr>
<th>PLAY</th>
<th>PLAY $\cap$ SHAKESPEARE</th>
<th>PLAY $\cap$ BASEBALL</th>
<th>PLAY $\cap$ CHILDREN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Play</td>
<td>Play</td>
<td>Play</td>
<td>children</td>
</tr>
<tr>
<td>Game</td>
<td></td>
<td>Play</td>
<td></td>
</tr>
<tr>
<td>Playing</td>
<td></td>
<td>Playing</td>
<td></td>
</tr>
<tr>
<td>Played</td>
<td></td>
<td>Played</td>
<td></td>
</tr>
<tr>
<td>Fun</td>
<td></td>
<td>Games</td>
<td></td>
</tr>
<tr>
<td>Games</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pat</td>
<td></td>
<td>Ball</td>
<td>children</td>
</tr>
<tr>
<td>Children</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ball</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Role</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plays</td>
<td>Plays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important</td>
<td>Important</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>Music</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part</td>
<td>Part</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run</td>
<td>Run</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friends</td>
<td></td>
<td></td>
<td>friends</td>
</tr>
<tr>
<td>Lot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td>Stage</td>
<td></td>
<td>toy</td>
</tr>
<tr>
<td>Toys</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audience</td>
<td>Audience</td>
<td></td>
<td>play</td>
</tr>
<tr>
<td>Theater</td>
<td></td>
<td></td>
<td>child</td>
</tr>
<tr>
<td>drama</td>
<td></td>
<td></td>
<td>school</td>
</tr>
<tr>
<td>scene</td>
<td></td>
<td></td>
<td>toys</td>
</tr>
</tbody>
</table>
The Construction of Meaning in ICA

- Distribution of semantic elements is modified: conditional probabilities
- Not all context is equal, but syntax structures context, determines how conditional probability distributions should be calculated
How to introduce syntactic structure?

- Dennis (2005): SP model - alignment via string edit theory
- Griffiths, Steyvers, Blei, & Tenenbaum (2004): topic model for semantics, hidden Markov model for syntax
- Solan, Horn, Ruppin, & Edelman (2005): ADIOS model infers syntactic patterns
How do children acquire language?

• **Tomasello** (2001): Young children’s early multiple-word productions are highly concrete, being based on particular words and phrases; grammatical development is word specific.

• Between 2 and 3 years, most verbs (87%) are used only in a single or very few forms. A few high-frequency verbs appear in several patterns:
  - *(Ann, Mommy, ...)* *draw*
    - *draw (man, doggie, ...)*
    - *draw (on paper, book, ... with pencil, crayons, ...)*
Bates & Goodman (2001):

- Fuzzy distinction between lexicon and grammar
- Syntax acquisition is a function of vocabulary size
  - At group level
  - At individual level
  - For both Italian and English
Psycholinguistics:
MacDonald, Perlmutter, & Seidenberg (1994)
Constraint-based lexicalist theory

Sentence processing as a form of ambiguity resolution:

• John SEMANTICS: animate, human, …
  THEMATIC ROLES: agent, experiencer, theme, goal
  LEXICAL CATEGORY: noun
  ARGUMENT STRUCTURE: (specifier) John (complement)

• cooked: SEMANTICS: prepare food, …
  LEXICAL CATEGORY: verb
  ARGUMENT STRUCTURE: (agent) cooked (complement)
  (agent) cooked (theme)
  (theme) cooked (complement)

John cooked
Lexicalized Tree-Adjoining Grammar

- Lexical items carry with them the syntactic structures in which they enter in the form of supertags.
- Two supertags for *with*:

```
  NP
   / \  
  NP  PP
     /  
    P   NP
       /  
      with
```

```
  VP
   / 
  PP
   / 
  P  NP
    /  
   with
```

“I saw the man in the park with a telescope”
Using supertags for constraint based syntactic analysis:

- Kim, Srinivas & Trueswell (2002)
- Syntactic sentence processing as a form of ambiguity resolution: possible supertags for the sentence “The police officer believed the victim was lying” (from Figure 1 of Kim et al. 2002)
QuickTime™ and its
TIFF (Tagged Image File Format) decompressor
are needed to see this picture.
Dependency tree of the sentence “Rolls-Royce said it expects its US sales to remain steady at about 1,200 cars” (after Yamada & Matsumoto, 2004; Nivre et al., 2007)
Propositional analysis after Kintsch (1998):

(1) SAID [ROLLS-ROYCE, 2]
(2) EXPECT [ROLLS-ROYCE, 3]
(3) REMAIN [SALES, STEADY]
(4) OF [ROLLS-ROYCE, SALES]
(5) US [SALES]
(6) AT [STEADY, 8]
(7) 1,200 [CARS]
(8) ABOUT [7]
Propositional structure mapped on the dependency tree:
Composable Dependency Units: Praful Mangalath

- The pretty girl saw a ghost

- girl  
  saw  
  ghost

- girl/pretty  
  saw/ghost  
  saw/girl

- Each vector is a probability distribution over semantic components
Comparing Tree Segments:

• Kullback-Leibler Divergence or Relative Entropy:
  \[ D(P \mid \mid Q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \]

• Earth Mover’s Distance
  \[ \sum_i \sum_j f_{ij} D_{ij} \]
  \[ \text{EMD}(P \mid \mid Q) = \frac{\sum_i \sum_j f_{ij} d_{ij}}{\sum_i \sum_j f_{ij}} \]
1: The president visited Ford's newly renovated automobile factory.
2: Ford manufactures cars.
3: The president manufactures cars.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>LSA cosine</th>
<th>ICA dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) The president visited Ford's newly renovated automobile factory.</td>
<td>0.51</td>
<td>0.77</td>
</tr>
<tr>
<td>2) Ford manufactures cars.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) The president manufactures cars.</td>
<td>0.84</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Sentence (1)

Dependency Units

factory:manufactures = 0.036
factory:cars = 0.094
unit similarity=0.94
Total similarity=$pWt$*unit similarity
=0.77

Sentence (2)

$pWt(ford|factory) = 0.76$
$pWt(ford|manufactures,cars) = 0.88$
$p W t = 0.82$
\[ p_{1Wt}(\text{president}|\text{visited, factory}) = 0.26 \]
\[ p_{2Wt}(\text{president}|\text{manufactures, cars}) = 0.04 \]
\[ p_{Wt} = 0.10 \]

factory: manufactures = 0.036 | factory: cars = 0.094
visited: manufactures = 0.028 | visited: cars = 0.087
unit similarity = 0.95
Total similarity = \( p_{Wt} \times \text{unit similarity} = 0.09 \)
• If we observe how children learn languages, we shall find that... people ordinarily show them the thing of which they would have them have the idea; and they repeat to them the name that stands for it, as “white”, “sweet”, “milk”, “sugar”, “cat”, “dog.”

John Locke
From Zwaan (in press):

• “Perceptual and motor activation occurs routinely during language comprehension”
• “If a theory based on perceptual and motor representations can account for all the findings that amodal, abstract and arbitrary symbol systems can account for, then why should we postulate the latter, given the wealth of evidence for the former and the lack of evidence for the latter?”
Levels of embedded representations:

- Symbolic representations
- Perceptual-motor representations
- Procedural representations
Noun advantage in early vocabulary acquisition:

Bates et al. (1995)
Identification of nouns and verbs through observational learning:

Gilette, Gleitman, Gleitman, & Lederer (1999)
Verb identification under three experimental conditions:
• Snedeker, Geren, & Shafto (2007): internationally adopted 6-year olds acquire vocabulary just like 2-year olds
Class cohesion as a function of learning (from Jones& Mewhort, 2007):
Nouns before verbs

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.
Class cohesion as a function of learning (from Jones & Mewhort, 2007):
Concrete before abstract
Thanks to

- Praful Mangalath
- National Science Foundation
- J. S. McDonnell Foundation