How Does the Brain Represent Word Meanings?

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 2010
Neurosemantics Research Team

Postdoctoral Fellow

Rob Mason

Research Scientists

Tom Mitchell
Marcel Just
Dean Pommerleau
Vladimir Cherkassky

PhD Students

Gustavo Sudre
Kai Min Chang
Leila Wehbe
Indra Rustandi
Mark Palatucci
Alona Fyshe
Typical stimuli

Each stimulus repeated several times

also spoken words
Functional MRI
fMRI activation for “bottle”:

Mean activation averaged over 60 different stimuli:

“bottle” minus mean activation:
Q1: Can one distinguish which word you’re thinking about based on fMRI?

Trained Classifier
(Bayes classifier, logistic regression, SVM, kNN, …)

(classifier as virtual sensor of mental state)
Training Classifiers over fMRI sequences

- Train the classifier function
  \[ \text{Mean}(\text{fMRI}(t+4), \ldots, \text{fMRI}(t+7)) \rightarrow \text{WordCategory} \]

- Preprocessing:
  - Adjust for head motion
  - Convert each image \( x \) to standard normal image

\[
x(i) \leftarrow \frac{x(i) - \mu_x}{\sigma_x}
\]

- Learning algorithms tried:
  - kNN (spatial correlation)
  - SVM
  - SVDM
  - Gaussian Naïve Bayes
  - Regularized Logistic regression \( \leftarrow \text{current favorite} \)
  - ...

- Feature selection methods tried:
  - Logistic regression weights, voxel stability, activity relative to fixation, regularization (L1, L2), ...

Classification task: is person viewing a “tool” or “building”?

- Classification accuracy for each participant from p4 to p1.
- All participants show a classification accuracy significantly above chance level (p<0.05).
Q2: Are neural representations similar across people?

Can we train on one group of people, decode for new person?
Local classifiers show where information is encoded

“tools” vs “buildings”

Accuracies of cubic 27-voxel classifiers centered at each voxel

[0.7-0.8]
Q3: Can we discover underlying principles of neural encodings?

arbitrary word $\xrightarrow{\text{Generative theory}}$ predicted brain activity

of word representation
Idea: Predict neural activity from corpus statistics of stimulus word

[Mitchell et al., Science, 2008]
<table>
<thead>
<tr>
<th>Semantic feature values: “<strong>celery</strong>”</th>
<th>Semantic feature values: “<strong>airplane</strong>”</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8368, eat</td>
<td>0.8673, ride</td>
</tr>
<tr>
<td>0.3461, taste</td>
<td>0.2891, see</td>
</tr>
<tr>
<td>0.3153, fill</td>
<td>0.2851, say</td>
</tr>
<tr>
<td>0.2430, see</td>
<td>0.1689, near</td>
</tr>
<tr>
<td>0.1145, clean</td>
<td>0.1228, open</td>
</tr>
<tr>
<td>0.0600, open</td>
<td>0.0883, hear</td>
</tr>
<tr>
<td>0.0586, smell</td>
<td>0.0771, run</td>
</tr>
<tr>
<td>0.0286, touch</td>
<td>0.0749, lift</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.0000, drive</td>
<td>0.0049, smell</td>
</tr>
<tr>
<td>0.0000, wear</td>
<td>0.0010, wear</td>
</tr>
<tr>
<td>0.0000, lift</td>
<td>0.0000, taste</td>
</tr>
<tr>
<td>0.0000, break</td>
<td>0.0000, rub</td>
</tr>
<tr>
<td>0.0000, ride</td>
<td>0.0000, manipulate</td>
</tr>
</tbody>
</table>
Predicted Activation is Sum of Feature Contributions

Predicted Celery = 0.84

\[ f_{eat}(celery) \]

from corpus statistics

\[ c_{14382, eat} \]

learned

\[ prediction_v = \sum_{i=1}^{25} f_i(w) c_{vi} \]

500,000 learned parameters

Predicted “Celery”
Predicted and observed fMRI images for “celery” and “airplane” after training on 58 other words.
Evaluating the Computational Model

• **Train** it using 58 of the 60 word stimuli
• **Apply** it to predict fMRI images for other 2 words
• **Test**: show it the observed images for the 2 held-out, and make it predict which is which

1770 test pairs in leave-2-out:
- Random guessing $\rightarrow$ 0.50 accuracy
- Accuracy above 0.61 is significant ($p<0.05$)

**Mean accuracy over 9 subjects**: 0.79
Participant P1

Eat

“Gustatory cortex”

Pars opercularis
(z=24mm)

Push

“sensory motor”

Postcentral gyrus
(z=30mm)

Run

“Biological motion”

Superior temporal sulcus (posterior)
(z=12mm)

Generative theory

“telephone”

Statistical features from a trillion-word text corpus

Mapping learned from fMRI data

predicted activity for “telephone”
Q4: What are the actual semantic primitives from which neural encodings are composed?

\[ v = \sum_{i=1}^{25} f_i(w) c_{vi} \]

25 verb co-occurrence counts??

predicted neural representation
## Alternative semantic feature sets

<table>
<thead>
<tr>
<th>PREDEFINED corpus features</th>
<th>Mean Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>486 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>50,000 word co-occurences</td>
<td>.76</td>
</tr>
<tr>
<td>300 Latent Semantic Analysis features</td>
<td>.73</td>
</tr>
<tr>
<td>50 corpus features from Collobert&amp;Weston ICML08</td>
<td>.78</td>
</tr>
</tbody>
</table>
### Alternative semantic feature sets

<table>
<thead>
<tr>
<th>PREDEFINED corpus features</th>
<th>Mean Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>486 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>50,000 word co-occurrences</td>
<td>.76</td>
</tr>
<tr>
<td>300 Latent Semantic Analysis features</td>
<td>.73</td>
</tr>
<tr>
<td>50 corpus features from Collobert&amp;Weston ICML08</td>
<td>.78</td>
</tr>
<tr>
<td><strong>218 features collected using Mechanical Turk</strong></td>
<td><strong>.83</strong></td>
</tr>
</tbody>
</table>

- Is it heavy?
- Is it flat?
- Is it curved?
- Is it colorful?
- Is it hollow?
- Is it smooth?
- Is it fast?
- Is it bigger than a car?
- Is it usually outside?
- Does it have corners?
- Does it have moving parts?
- Does it have seeds?
- Can it break?
- Can it swim?
- Can it change shape?
- Can you sit on it?
- Can you pick it up?
- Could you fit inside of it?
- Does it roll?
- Does it use electricity?
- Does it make a sound?
- Does it have a backbone?
- Does it have roots?
- Do you love it?
- …

features authored by Dean Pomerleau.

feature values 1 to 5

features collected from at least three people

people provided by Amazon’s “Mechanical Turk”
## Alternative semantic feature sets

<table>
<thead>
<tr>
<th>PREDEFINED corpus features</th>
<th>Mean Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>486 verb co-occurrences</td>
<td>.79</td>
</tr>
<tr>
<td>50,000 word co-occurences</td>
<td>.76</td>
</tr>
<tr>
<td>300 Latent Semantic Analysis features</td>
<td>.73</td>
</tr>
<tr>
<td>50 corpus features from Collobert &amp; Weston ICML08</td>
<td>.78</td>
</tr>
<tr>
<td><strong>218 features collected using Mechanical Turk</strong></td>
<td><strong>.83</strong></td>
</tr>
<tr>
<td><strong>20 features discovered from the data</strong></td>
<td><strong>.87</strong></td>
</tr>
</tbody>
</table>
Discovering shared semantic basis

[Rustandi et al., 2009]

* trained using Canonical Correlation Analysis
Multi-study (WP+WO) Multi-subject (9+11) CCA
Top Stimulus Words

<table>
<thead>
<tr>
<th>most positive stimuli</th>
<th>component 1</th>
<th>component 2</th>
<th>component 3</th>
<th>component 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>apartment church</td>
<td>screwdriver</td>
<td>telephone</td>
<td>pants</td>
<td></td>
</tr>
<tr>
<td>closet barn</td>
<td>pliers refrigerator</td>
<td>butterfly</td>
<td>dress</td>
<td></td>
</tr>
<tr>
<td>house</td>
<td>knife hammer</td>
<td>bicycle beetle</td>
<td>glass</td>
<td></td>
</tr>
<tr>
<td>barn</td>
<td></td>
<td>dog</td>
<td>coat</td>
<td></td>
</tr>
</tbody>
</table>

shelter? manipulation? things that touch me?
Additional Directions

• Model for abstract words (love, justice, anxiety,...)
  – preliminary: accuracies similar to those for concrete nouns

• Model phrases (“firm tomato”)
  – [Chang et al., ACL2009]: composing corpus statistics for
    <adjective> and <noun> predicts fMRI for <adjective noun>

• MEG imaging (1 msec time resolution)
  – preliminary results: can train classifiers to detect both where
    and when neural activity codes word meanings, and stimulus
    percepts

• ML algorithms that build cumulative models from
  many (100’s of) data sets
Where Next?

• What will a “theory” of the brain (or the cell) look like?

• Set of architectural organizing principles,
  • and a detailed computational model that follows them

• How will we learn it?

• Current approaches are data-starved
  • Need algorithms that learn cumulatively from
    – many experiments
    – priors gleaned from research literature
    – priors that express researcher’s hypotheses
    – optimal planning of next experiment
thank you!