Decoding the Information Encoded in Neural Activity

Tom M. Mitchell
Carnegie Mellon University
April 2017
How does neural activity encode word meanings?
How does neural activity encode word meanings?

How does brain combine word meanings into sentence meanings?
Neurosemantics Research Team

Research Scientists

- Erika Laing
- Tom Mitchell
- Marcel Just

- Leila Wehbe
- Dan Schwartz
- Alona Fyshe
- Mariya Toneva
- Mark Palatucci
- Gustavo Sudre
- Nicole Rafidi

Recent/Current PhD Students

funding: NSF, NIH, IARPA, Keck
Functional MRI
Typical stimuli
fMRI activation for “bottle”:

Mean activation averaged over 60 different stimuli:

“bottle” minus mean activation:
Classifiers trained to decode the stimulus word

Trained Classifier

(SVM, Logistic regression, Deep net, Bayesian classifier ...)

(classifier as virtual sensor of mental state)
Classification task: is person viewing a “tool” or “building”?

Participants

Classification accuracy

statistically significant

p<0.05
Are neural representations similar across people?

Can we train classifiers on one group of people, then decode from new person?
Are representations similar across people?

YES

classify which of 60 items
Lessons from fMRI Word Classification

Neural representations similar across
• people
• language
• word vs. picture

Easier to decode:
• concrete nouns
• emotion nouns

Harder to decode:
• abstract nouns
• verbs*

* except when placed in context
Predictive Model?

Arbitrary noun ➔ Predicted fMRI activity
Predictive Model?

[Mitchell et al., *Science*, 2008]

Input noun: “telephone”

Retrieve text statistics

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

Predicted fMRI activity

vector representing word meaning

trillion word text collection

trained on other fMRI data

[Mitchell et al., *Science*, 2008]
Represent stimulus noun by co-occurrences with 25 verbs*

<table>
<thead>
<tr>
<th>Semantic feature values: “celery”</th>
<th>Semantic feature values: “airplane”</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>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>

* in a trillion word text collection
Predicted Activation is Sum of Feature Contributions

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

500,000 learned \( c_{vi} \) parameters

\( f_{\text{eat}}(\text{celery}) \) from corpus statistics

\( c_{14382,\text{eat}} \) learned

Celery = 0.84 + 0.35 + 0.32 + \( \ldots \)

Predicted “Celery”

high

low
Predicted and observed fMRI images for “celery” and “airplane” after training on other nouns. [Mitchell et al., Science, 2008]
Evaluating the Computational Model

- Leave two words out during training

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
Learned activities associated with meaning components

Participant P1

Semantic feature:

- **Eat** "Gustatory cortex"  
  Pars opercularis (z=24mm)

- **Push** "somato-sensory"  
  Postcentral gyrus (z=30mm)

- **Run** "Biological motion"  
  Superior temporal sulcus (posterior) (z=12mm)
## 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>
</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-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>218 features collected using <em>Mechanical Turk</em></td>
<td>.83</td>
</tr>
<tr>
<td>20 features discovered from the data**</td>
<td>.86</td>
</tr>
</tbody>
</table>

* developed by Dean Pommerleau
** developed by Indra Rustandi
Discovering shared semantic basis

1. Use CCA to discover latent features across subjects

CCA abstraction
\[ f_k(w) = \sum_v x_v c_{vi} \]

specific to study/subject

- subj 1, word+pict
- subj 9, word+pict
- subj 10, word only
- subj 20, word only

20 learned latent features

[Rustandi et al., 2009]
Canonical correlation analysis

\[ \text{Corr}(A, B) = \frac{1}{N} \sum_{i=1}^{N} \frac{(A_i - \bar{A})}{\sigma_A} \frac{(B_i - \bar{B})}{\sigma_B} \]

Each column is one fMRI image

maximally correlated

\[ w'_X X \quad w'_Y Y \]

[slide courtesy of Indra Rustandi]
Discovering shared semantic basis

1. Use CCA to discover latent features

\[
f_k(w) = \sum_v x_v c_{vi}
\]

specific to study/subject

- subj 1, word+pict
- subj 9, word+pict
- subj 10, word only
- subj 20, word only

20 learned latent features

[Rustandi et al., 2009]
Discovering shared semantic basis

1. Use CCA to discover latent features
2. Train regression to predict them

specific to study/subject

CCA abstraction
\[ f_k(w) = \sum_v x_v c_{v} \]

[\text{Rustandi et al., 2009}]

independent of study/subject

218 MTurk features

20 learned latent features

word \(w\)
Discovering shared semantic basis

1. Use CCA to discover latent features
2. Train regression to predict them
3. Invert CCA mapping

[Rustandi et al., 2009]
CCA Components: Top Stimulus Words

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

shelter? manipulation? things that touch my body?
Timing?
MEG: Stimulus “hand” (word plus line drawing)

[Sudre et al., *Neurolmage* 2012]
100 ms word length right diagonalness

[Sudre et al., under review]

50 ms

0 800 ms

[Sudre et al., NeuroImage 2012]
100 ms

[Sudre et al., 2012]
150 ms

aspect ratio

word length
internal details

[Sudre et al., 2012]
IS IT HAIRY?

aspect ratio

internal details

internal details

[Sudre et al., 2012]
IS IT HOLLOW?

IS IT MADE OF WOOD?

white pixel count horizontalness

IS IT HAIRY?
IS IT AN ANIMAL?

250 ms

[Sudre et al., 2012]
CAN YOU PICK IT UP?
CAN YOU HOLD IT?
IS IT BIGGER THAN A CAR?

IS IT MAN-MADE?
IS IT ALIVE?
CAN IT BITE OR STING?

WAS IT EVER ALIVE?
DOES IT GROW?

IS IT ALIVE?
DOES IT GROW?

0-800 ms

[Sudre et al., 2012]
COULD YOU FIT INSIDE IT? DOES IT HAVE FOUR LEGS?

CAN YOU PICK IT UP? CAN YOU HOLD IT?

CAN YOU HOLD IT IN ONE HAND?

IS IT MAN-MADE? WAS IT EVER ALIVE?

IS IT ALIVE? CAN IT BEND?

350 ms

[Sudre et al., 2012]
CAN YOU PICK IT UP?
IS IT BIGGER THAN A CAR?
DOES IT HAVE CORNERS?
CAN YOU PICK IT UP?
IS IT TALLER THAN A PERSON?
IS IT MAN-MADE?
WAS IT EVER ALIVE?
WAS IT INVENTED?
IS IT MANUFACTURED?
DOES IT HAVE FEELINGS?
IS IT ALIVE?
CAN YOU HOLD IT?

IS IT ALIVE?

IS IT MANUFACTURED?

IS IT HOLLOW?

IS IT HOLLOW?

DOES IT GROW?

IS IT BIGGER THAN A BED?

WAS IT INVENTED?

[Cudre et al., under review]

[Sudre et al., 2012]
IS IT TALLER THAN A PERSON?
CAN YOU PICK IT UP?

DOES IT GROW?

CAN YOU PICK IT UP?

IS IT BIGGER THAN A BED?

CAN YOU HOLD IT IN ONE HAND?

[Sudre et al., under review]

[Sudre et al., 2012]
CAN IT BE EASILY MOVED?

IS IT ALIVE?
IS IT MAN-MADE?
WAS IT EVER ALIVE?

[Sudre et al., under review]

[Sudre et al., 2012]
Details
Color: decodability* of feature “wordlength” (peak decodability 100-150 msec)

* % of feature variance predicted by MEG, mean across 9 subjects
Color = decodability of "grasping" features (initial peak: 200-300 msec)

[Sudre et al., 2012]
Color= decodability of “animacy” features (initial peak: 150-200 msec)

[Sudre et al., 2012]
20 most accurately decoded semantic features out of 218

[G. Sudre et al., 2012]
Sentence Reading

Nicole Rafidi, Erika Liang, Dan Schwartz

How does brain parse sentences?

Which word meanings are activated when, as the sentence is being understood?
student
found
the
hammer.
32 sentences x 15 repetitions:

- **16 Active**: “A student found the hammer”
- **16 Passive**: “The hammer was found by a student”
Sentence MEG activity:

| A     | student | found | the    | hammer | +  |

![MEG activity graph]

**brain locations / MEG sensors**

**time →**
**Sentence MEG activity:**

<table>
<thead>
<tr>
<th>A</th>
<th>student</th>
<th>found</th>
<th>the</th>
<th>hammer</th>
<th>+</th>
</tr>
</thead>
</table>

when is the brain ‘thinking about’ each earlier word? (e.g., first noun, ‘student’?)
Sentence MEG activity:

| A | student | found | the | hammer | + |

when is the brain ‘thinking about’ each earlier word? (e.g., first noun, ‘student’?)

when can we decode first noun (“student”) from sliding 100 msec window of MEG?

100 msec =
When does brain activate meaning of **first noun** in sentence?  
[Nicole Rafidi et al., HBM 2015]
thank you!