Grasping the finer point: Metaphor identification in text and brain imaging data

Ekaterina Shutova

ILLC
University of Amsterdam
Stanford University, 23 August 2018
What is metaphor?
How does metaphor work?

Association between two concepts
(Gentner, 1983; Lakoff and Johnson, 1980)

**POLITICALSYSTEM** is a **MECHANISM**

\[ \text{target} \quad \text{source} \]

“rebuilding the campaign machinery”
“Time to mend our foreign policy”
“20 Steps towards a working democracy”
Today's talk

1. Metaphor identification method
   (Rei, Bulat, Kiela & Shutova, EMNLP 2017)

2. Using NLP techniques to study metaphor processing in the brain
   (Gamez-Gjokic, Maillard, Bulat & Shutova, forthcoming)
Metaphor identification: Existing approaches

Linguistic resources:
- Semantic roles
  (Gedigian et al., 2006)
- Concreteness
  (Turney et al., 2011)
- Imageability
  (Strzalkowski et al., 2013)
- WordNet supersenses
  (Tsvetkov et al., 2014)

Data-driven methods & cognitive features:
- Clustering with sparse distributional features
  (Shutova et al., 2010)
- Visual vectors
  (Shutova et al., 2016)
- Attribute-based vectors
  (Bulat et al., 2017)
A neural architecture for metaphor processing

Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection.
Rei, Bulat, Kiela & Shutova, EMNLP 2017.

- Supervised classification setting
- Identifying metaphorical uses of verbs and adjectives

<table>
<thead>
<tr>
<th>Mohammad et al. (2016)</th>
<th>Tsvetkov et al. (2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verb noun</strong></td>
<td><strong>Adj. noun</strong></td>
</tr>
<tr>
<td>boost economy</td>
<td>cloudy future</td>
</tr>
<tr>
<td>boost voltage</td>
<td>cloudy sky</td>
</tr>
<tr>
<td><strong>Class</strong></td>
<td><strong>Class</strong></td>
</tr>
<tr>
<td>met.</td>
<td>met.</td>
</tr>
<tr>
<td>lit.</td>
<td>lit.</td>
</tr>
</tbody>
</table>

Metaphor identification in text and brain imaging data
Approach

**INPUT**: skip-gram word embeddings
- 100-dimensional
- trained on Wikipedia

**OUTPUT**: a metaphoricity score between 0 and 1
Approach

**INPUT:** skip-gram word embeddings
- 100-dimensional
- trained on Wikipedia

**OUTPUT:** a metaphoricity score between 0 and 1

**Key intuitions:**
1. model domain interaction via **gating**
2. specialise word representations
3. quantify metaphoricity via a **weighted similarity** function
Word representation gating

\[ g = \sigma(W_g x_1) \]
\[ \tilde{x}_2 = x_2 \odot g \]

- \( W_g \) — a weight matrix
- \( \sigma \) — sigmoid activation function
- \( \odot \) — element-wise multiplication.

Some properties of the source domain are projected onto the target.
Specialisation

\[ z_1 = \tanh(W_{z_1} x_1) \]
\[ z_2 = \tanh(W_{z_2} \tilde{x}_2) \]

Vector Space Mapping

The original method uses basic pre-trained skip-gram vectors. Let’s add a transformation that maps them into a space that is specific for metaphor detection. Importantly, we will use separate mappings for adjectives and nouns. The weight matrices are optimised during training, while the pre-trained embeddings are kept fixed.
Weighted similarity

If the input vectors $x_1$ and $x_2$ are normalised to unit length, the cosine similarity between them is equal to their dot product:

$$\cos(x_1, x_2) \propto \sum_i x_{1,i}x_{2,i}$$

We can formulate this as a small neural network:
We can instead create a version where vector \( m \) is passed through another layer, with weights that are optimised during training.

\[
m_i = z_1,i z_2,i
\]

\[
d = \gamma(W_d m)
\]
Supervised similarity network

The final network architecture, using:

- Word representation gating
- Specialisation
- Vector combination based on weighted cosine
## Results: Adjectives

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsvetkov et al. (2014)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td><strong>85</strong></td>
</tr>
<tr>
<td>Shutova et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>linguistic</td>
<td>-</td>
<td>73</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td>multimodal</td>
<td>-</td>
<td>67</td>
<td><strong>96</strong></td>
<td>79</td>
</tr>
<tr>
<td>Bulat et al. (2017)</td>
<td>-</td>
<td>85</td>
<td>71</td>
<td>77</td>
</tr>
<tr>
<td>FFN skip-gram</td>
<td><strong>77.6</strong></td>
<td><strong>86.6</strong></td>
<td><strong>65.4</strong></td>
<td><strong>74.4</strong></td>
</tr>
<tr>
<td>SSN skip-gram</td>
<td><strong>82.2</strong></td>
<td><strong>91.1</strong></td>
<td><strong>71.6</strong></td>
<td><strong>80.1</strong></td>
</tr>
</tbody>
</table>
## Results: Verbs

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shutova et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>linguistic</td>
<td>-</td>
<td>67</td>
<td>76</td>
<td>71</td>
</tr>
<tr>
<td>multimodal</td>
<td>-</td>
<td>65</td>
<td><strong>87</strong></td>
<td><strong>75</strong></td>
</tr>
<tr>
<td>FFN skip-gram</td>
<td>71.2</td>
<td>70.4</td>
<td>71.8</td>
<td>70.5</td>
</tr>
<tr>
<td>SSN skip-gram</td>
<td><strong>74.8</strong></td>
<td><strong>73.6</strong></td>
<td><strong>76.1</strong></td>
<td><strong>74.2</strong></td>
</tr>
</tbody>
</table>
Qualitative analysis

Metaphor identification in text and brain imaging data
Applications in social science (and beyond)

- Metaphor as a predictor of influence/popularity of politicians
- Vinod, Dan and I
- Facebook dataset
- The number of metaphors used can serve as a predictor of the number of shares, likes etc.
- Looking at the identity of the metaphors next
Decoding literal and metaphorical sentences in the brain

Can we use semantic models to better understand metaphor processing in the brain?
Gamez-Djokic, Maillard, Bulat and Shutova.

Experiments with brain imaging data

- **Data**: fMRI neural activation patterns associated with the meaning of literal and metaphorical sentences (Gamez-Djokic et al, forthcoming)
- Verbs in their metaphorical and literal contexts
- **Task**: decode patterns of brain activity
- Using data-driven semantic models
functional Magnetic Resonance Imaging (fMRI)

- **Voxel**: a 3x3x6mm$^3$ cube of brain tissue
- **Voxel value**: intensity of brain activity in that voxel
- **fMRI image**: vector of voxel values (represents brain activation pattern)
Our brain imaging dataset

- 15 participants
- 31 unique hand-action verbs
- 200 sentences
- 5 conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affirmative Literal</td>
<td>She’s <em>grasping</em> the cookie</td>
</tr>
<tr>
<td>Affirmative Metaphor</td>
<td>She’s <em>grasping</em> the lecture</td>
</tr>
<tr>
<td>Negated Literal</td>
<td>He’s not <em>grasping</em> the bill</td>
</tr>
<tr>
<td>Negated Metaphor</td>
<td>He’s not <em>grasping</em> the problem</td>
</tr>
<tr>
<td>Affirmative Paraphrase</td>
<td>She’s understanding the lecture</td>
</tr>
</tbody>
</table>
Stimuli presentation

- Disambiguation – object:
  
  *The physics lecture (2 seconds)*

- Interval:
  
  *(0.5 seconds)*

- Stimulus:
  
  *She is grasping the lecture (6 seconds)*

- Rest:
  
  *(8 seconds)*
Semantic models

1. **Linguistic models**
   - Word representations
   - Compositional models

2. **Visually grounded models**
   - word and phrase representations
   - learned from images

3. **Multimodal models**
   - combining linguistic and visual information
Linguistic models

- Individual words: \textit{Verb} and \textit{Object}
  - GloVe word embeddings (Pennington et al. 2014)
- \textit{VerbObject}: concatenation of verb and object embeddings
- \textit{Addition}: addition of verb and object embeddings
- \textit{LSTM}: learn representations for verb-object phrases
  - trained on the natural language inference task
  - taking Glove word embeddings as input
Visual representations

1. Retrieve images for a word or phrase using Google Search
2. **Transfer learning** to extract image embeddings:
   - Convolutional neural network trained on the ImageNet classification task (Kiela and Bottou, 2014)

   - Forward pass
   - Use penultimate layer (FC7) as image embedding

### Diagram
- ImageNet synsets
- Convolutional layers: C1-C2-C3-C4-C5
- Fully-connected layers: FC6, FC7, FC8
- Imagenet labels:
  - African elephant
  - Wall clock
  - ...
Visual and multimodal models

Visual models:

- Individual words: *VERB* and *OBJECT*
- *VERB*OBJECT*: concatenation of verb and object embeddings
- **ADDITION**: addition of verb and object embeddings
- **PHRASE**: visual representation for the whole phrase

Multimodal models:

- Concatenation of the respective linguistic and visual models
- with the exception of LSTM
Decoding brain activity

Similarity-based decoding (Anderson et al., 2016)

Slide credit: Andrew Anderson
systematically biased the fMRI correlation structure (calculated next) to look like that of the semantic model, and can influence how word labels are as-

Figure 2: Similarity-decoding algorithm (adapted from Anderson et al. 2016).

Fisher’s r to z transform (arctanh) is typically used because these values could reveal the correct answer to decoding. The two model similarity vectors were then compared to the two fMRI similarity vectors, resulting in four correlation matrices. Entries corresponding to the two test words were drawn from both the model and fMRI similarity matrices. 

Similarity vectors for the two test words were formed using Fisher’s r to z (arctanh). If the sum of z-transformed correlations between the correctly matched pair exceeded the sum of correlations for the incongruent pair, decoding was scored a success, otherwise a failure. This process was then repeated for all word pairs, with the mean accuracy of all test pairs giving a final measure of success.

Meta analysis coefficient. In the similarity-decoding method, performance is critical to the procedure. z noticeably differs from r in both model and fMRI similarity vectors were represented. Similarity vectors for the two test words were formed using Fisher’s r to z (arctanh). If the sum of z-transformed correlations between the correctly matched pair exceeded the sum of correlations for the incongruent pair, decoding was scored a success, otherwise a failure. This process was then repeated for all word pairs, with the mean accuracy of all test pairs giving a final measure of success.

Later undertaken in Section 4.1)

A single representation of each word was built by taking the voxel-wise mean of all five presentations giving a final measure of success.

Later in this section, there were no significant differences in decoding accuracy arising from tests using 68/70 versus 70 words.

However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, and consequently biased decoding performance.

However, under most circumstances r to z is not optimised a mapping between fMRI data and semantic model, and consequently biased decoding performance.

\[
\text{Brain similarity matrix}
\]

\[
\text{Model similarity matrix}
\]

\[
\text{Brain similarity vectors}
\]

\[
\text{Model similarity vectors}
\]

**Slide credit:** Andrew Anderson
Similarity-based decoding systematically biased the fMRI correlation structure (calculated next) to look like that of the semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, it is not prone to modelling and decoding fMRI noise as in classic cases of double dipping (Kriegeskorte et al., 2009). Indeed, as we report later in this section, there were no significant differences in decoding accuracy arising from tests using voxel selection on 68/70 versus 70 words.

A single representation of each word was built by taking the voxel-wise mean of all five presentations of the word for the 500 selected voxels. An fMRI similarity matrix for all 70 words was then calculated. Similarity vectors for the two test words were drawn from both the model and fMRI similarity matrices. Entries corresponding to the two test words in both model and fMRI similarity vectors were removed because these values could reveal the correct answer to decoding. The two model similarity vectors were then compared to the two fMRI similarity vectors by correlation, resulting in four correlation values. These correlation values were transformed using Fisher’s r to z (arctanh). If the sum of z-transformed correlations between the correctly matched pair exceeded the sum of correlations for the incongruent pair, decoding was scored a success, otherwise a failure. This process was then repeated for all word pairs, with the mean accuracy of all test iterations giving a final measure of success.

Fisher’s r to z transform (arctanh) is typically used to test for differences between correlation coefficients. It transforms the correlation coefficient r to a value z, where z has amplified values at the tails of the correlation coefficient (r otherwise ranges between -1 and 1). This is to make the sampling distribution of z normally distributed, with approximately constant variance values across the population correlation coefficient. In the similarity-decoding method used here, z is evaluated in decoding because it is a more principled metric to compare and combine (as later undertaken in Section 4.1).

However, under most circumstances r to z is not critical to the procedure. z noticeably differs from r only when correlations exceed .5, and r to z changes decoding behaviour in select circumstances. Specifically r to z can influence how word labels are assigned.

Slide credit: Andrew Anderson
Similarity-based decoding systematically biased the fMRI correlation structure (calculated next) to look like that of the semantic model, and consequently biased decoding performance. However, as similarity-based decoding does not optimise a mapping between fMRI data and semantic model, it is not prone to modelling and decoding fMRI noise as in classic cases of double dipping (Kriegeskorte et al., 2009). Indeed, as we report later in this section, there were no significant differences in decoding accuracy arising from tests using voxel selection on 68/70 versus 70 words.

A single representation of each word was built by taking the voxel-wise mean of all five presentations of the word for the 500 selected voxels. An fMRI similarity matrix for all 70 words was then calculated. Similarity vectors for the two test words were drawn from both the model and fMRI similarity matrices. Entries corresponding to the two test words in both model and fMRI similarity vectors were removed because these values could reveal the correct answer to decoding. The two model similarity vectors were then compared to the two fMRI similarity vectors by correlation, resulting in four correlation values. These correlation values were transformed using Fisher's r to z (arctanh). If the sum of z-transformed correlations between the correctly matched pair exceeded the sum of correlations for the incongruent pair, decoding was scored a success, otherwise a failure. This process was then repeated for all word pairs, with the mean accuracy of all test iterations giving a final measure of success.

Fisher's r to z transform (arctanh) is typically used to test for differences between correlation coefficients. It transforms the correlation coefficient r to a value z, where z has amplified values at the tails of the correlation coefficient (r otherwise ranges between -1 and 1). This is to make the sampling distribution of z normally distributed, with approximately constant variance values across the population correlation coefficient. In the similarity-decoding method used here, z is evaluated in decoding because it is a more principled metric to compare and combine (as later undertaken in Section 4.1).

However, under most circumstances r to z is not critical to the procedure. z noticeably differs from r only when correlations exceed .5, and r to z changes decoding behaviour in select circumstances. Specifically r to z can influence how word labels are as-

\[
\text{Decoding:}
\]

\[
\text{if } \text{corr}(A,3)+\text{corr}(B,6)>\text{corr}(A,6)+\text{corr}(B,3) \quad A=3; \quad B=6;
\]

\[
\text{else} \quad A=6; \quad B=3;
\]

Slide credit: Andrew Anderson
Results: Linguistic models

- The models were evaluated in terms of **decoding accuracy**
- Significance was determined via permutation testing

<table>
<thead>
<tr>
<th></th>
<th>Literal</th>
<th>Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT</td>
<td>0.51</td>
<td>0.67</td>
</tr>
<tr>
<td>VERB</td>
<td>0.71</td>
<td>0.54</td>
</tr>
<tr>
<td>VERBOBJECT</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>ADDITION</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.6</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Metaphor identification in text and brain imaging data
Results: Visual and multimodal models

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Literal</th>
<th>Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISUAL OBJECT</td>
<td>0.58</td>
<td>0.44</td>
</tr>
<tr>
<td>VISUAL VERB</td>
<td>0.47</td>
<td>0.66</td>
</tr>
<tr>
<td>VISUAL VERBOBJECT</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>VISUAL ADDITION</td>
<td>0.47</td>
<td>0.68</td>
</tr>
<tr>
<td>VISUAL PHRASE</td>
<td>0.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>
## Results: Visual and multimodal models

<table>
<thead>
<tr>
<th></th>
<th>Literal</th>
<th>Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VISUAL OBJECT</strong></td>
<td>0.58</td>
<td>0.44</td>
</tr>
<tr>
<td><strong>VISUAL VERB</strong></td>
<td>0.47</td>
<td><strong>0.66</strong></td>
</tr>
<tr>
<td><strong>VISUAL VERBOBJECT</strong></td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>VISUAL ADDITION</strong></td>
<td>0.47</td>
<td><strong>0.68</strong></td>
</tr>
<tr>
<td><strong>VISUAL PHRASE</strong></td>
<td>0.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Literal</th>
<th>Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MULTIMODAL OBJECT</strong></td>
<td>0.62</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>MULTIMODAL VERB</strong></td>
<td>0.52</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td><strong>MULTIMODAL VERBOBJECT</strong></td>
<td>0.48</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>MULTIMODAL ADDITION</strong></td>
<td>0.55</td>
<td><strong>0.72</strong></td>
</tr>
</tbody>
</table>
What can we learn from this?

1. The verb embedding is successful in decoding brain activity in the literal case.

2. The object embedding and compositional models are more successful in the metaphor case.

   This may suggest that humans pay more attention to the object when interpreting metaphor (speculation).

3. Visual representations yield significant decoding accuracies in the metaphor case, but not literal.

   This suggests that the visual information plays a role in metaphor processing (speculation).
I would like to thank...

**Collaborators:** Marek Rei, Luana Bulat, Douwe Kiela, Jean Maillard and Vesna Gamez-Djokic

- Stanford NLP group for the invitation!