Semantic Vectors as a Tool for Psychological Investigation

Eyal Sagi
eya@u.northwestern.edu

Premise

**Text as a proxy for its author**
The psychological makeup of a person influences their use of language. It is possible to identify these influences and use them to test psychological theories.

Overview

**Words represented as vectors in a semantic space**
- **Data points:** Word spans
- **Identified and labeled based on theory**
- **Represented by sum of word vectors**

- **Basic measure:** Angle between vectors (i.e. distance)

Semantic Spaces

**Topicality:** Words that occur together are likely to relate to the same topic
- Based on *Infomap* (Takayama et al. 1999)
- Similar to *Latent Semantic Analysis* (Deerwester et al., 1990)

- **Semantic space is based on patterns of word co-occurrence**
- **Linear space generated using Singular Value Decomposition (generalized factoring)**

Convergence in language use

**Reaching agreement**

(3-party negotiations; ~0.5M words)
- Intra-Coalition
- Extra-Coalition

Mean Similarity of Use

<table>
<thead>
<tr>
<th>Quarter of the Negotiation</th>
<th>Conserv = 1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberals</td>
<td>0.25</td>
<td>0.15</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Conservatives</td>
<td>0.15</td>
<td>0.20</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Error bars in graphs represent standard error of the mean.

Moral Rhetoric

**Federal Shutdown** (Tweeter; ~9.5M words)

Analysis based on Moral Foundations Theory (Haidt & Joseph 2004)

- **Liberals**
  - During Shutdown
  - After Shutdown
- **Conservatives**
  - During Shutdown
  - After Shutdown

Retweeting Behavior

<table>
<thead>
<tr>
<th>Moral Loading</th>
<th>Authority</th>
<th>Harm</th>
<th>Loyalty</th>
<th>Purity</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberals</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Conservatives</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Error bars in graphs represent standard error of the mean.

Framing effects

**Framing of ‘terror’ over time**

(Senate speeches; 180M words)

Higher numbers indicate a greater degree of framing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberals</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Conservatives</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Source: Sagi (2010)

Error bars in graphs represent standard error of the mean, number of words at bottom of bars.

Nouns vs. Verbs

(Project Gutenberg; 240M words)

Average angle for word context vectors by grammatical category.

References