Interpretability

Advanced Machine Learning for NLP
Jordan Boyd-Graber
NEED FOR INTERPRETABILITY

Slides adapted from Tobias Schnabel
Evaluating Word Embeddings

• Collected without reference to embeddings
  ○ Balance rare and frequent words (e.g., play vs. devour)
  ○ Balance POS classes (e.g., skillfully vs. piano)
  ○ Balance abstractness/concreteness (e.g., eagerness vs. table)

• See if embeddings can answer questions

• Perhaps not right questions to distinguish methods
Setup

- **Embeddings**
  - Prediction-based: CBOW and Collobert&Weston (CW)
  - Reconstruction-based: CCA, Hellinger PCA, Random Projections, GloVe
  - Trained on Wikipedia (2008), made vocabularies the same

- **Details**
  - Options came from position $k = 1, 5, 50$ in NN from each embedding
  - 100 query words x 3 ranks = 300 subtasks
  - Users of Amazon Mechanical Turk answered 50 such questions

- **Win score**: Fraction of votes for each embedding, averaged
Winners

![Bar chart showing performance of different models](chart.png)

- **1 NN**
- **5 NN**
- **50 NN**

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand. Proj</td>
<td>0.10</td>
</tr>
<tr>
<td>H-PCA</td>
<td>0.15</td>
</tr>
<tr>
<td>C&amp;W</td>
<td>0.20</td>
</tr>
<tr>
<td>TSCCA</td>
<td>0.25</td>
</tr>
<tr>
<td>GloVe</td>
<td>0.30</td>
</tr>
<tr>
<td>CBOW</td>
<td>0.35</td>
</tr>
</tbody>
</table>
What about Intruders?

- Query word
- Nearest neighbors
- Coherent
- Intruder
What about Intruders?

<table>
<thead>
<tr>
<th>(a) finally</th>
<th>(b) eventually</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c) put</td>
<td>(d) immediately</td>
</tr>
</tbody>
</table>
What about Intruders?

(a) Normalized scores by global word frequency.
Downstream Tasks?

Intrinsic performance

Extrinsic performance

Normalized scores by global word frequency.

F1 chunking results