Vector-based Semantics

Credit: much of this lecture is based on Chris Potts' excellent videos for Stanford's CS224u class

Agenda

Review of vector-based representations and similarity metrics

Evaluating the goodness of our vector representations using psycholinguistics data

What are word vectors good for (and is it really "semantics")?
- Find the most similar words
- Solve word relation puzzles like analogies
- Compose their meanings (??)

What semantics problems are difficult to do with word vectors?
- negation/antonyms/exclusion
- inference
- composition of meaning for sentences

user → document/movie
word → document matrix

\[ d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5 \quad \ldots \quad d_N \]

abandon
abdicate
abjure
academic
zygodactyl
zyurgy

What is a document
- Shakespeare's plays
- Wikipedia articles
- IMDB reviews for a movie
- speeches by politicians
- real estate listings

Compare the similarity of "documents"
- author attribution/plagiarism detection/document de-duplication
- clustering into categories
- recommendation systems
- compare documents by "tone"
- \( \text{doc} = \text{query} \)

Comparing rows
- what words are similar to each other
Zellig Harris (1954) distributional statements can cover all of the material of a language without requiring support from other types of information.

Turney and Pantel (2010) “from freq to meaning” if units of text have similar vectors in a text frequency matrix than they tend to have similar meanings.

```
Word x word matrix / context
```

```
abandon ablolute abscend abscond ...
```

Define a co-occurrence matrix

```
V x V
```

take $5k$ most freq.

```
word 2 vec
```

Zipfian distribution:

```
<table>
<thead>
<tr>
<th>The of a form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
</tr>
<tr>
<td>Zymurgy</td>
</tr>
<tr>
<td># of words w/ freq</td>
</tr>
</tbody>
</table>
```

Aside:
co-occur in a document
sentence
window of words
more complex contexts
dependency pattern

The pictures are beautiful.
The pictures of the old man holding his first grandchild are beautiful.

Lin and Pantel (2001) DRT

duty and responsibility

modified by similar sets of adjectives
additional - amod
administrative - amod
congressional - amod
constitutional - amod

objects of similar sets of verbs:
abandon - obj-of
dedicate - obj-of
assign - obj-of
assume -
breach -
pair pattern co-occurrence matrix

\[ x \text{ works with } y \]
\[ x \text{ cuts } y \]

\langle \text{mason, stone} \rangle
\langle \text{carpenter, wood} \rangle
\langle \text{scissors, paper} \rangle

Vector Comparison

\begin{array}{c|c|c}
\text{words} & \text{doc} x & \text{doc} y \\
\hline
\text{A} & 2 & 4 \\
\text{B} & 10 & 15 \\
\text{C} & 14 & 10 \\
\end{array}

\text{Euclidean distance}
\text{vectors } u, v \text{ of dimension } N

\[ \sqrt{\sum_{i=1}^{N} (u_i - v_i)^2} \]

\text{Goal: related words are close, unrelated words are far apart}

\text{Vector L2 (length) normalization}
\text{normalization of } u \text{ is a vector } \hat{u} \text{ obtained by dividing each element of } u \text{ by } \|u\| \text{ (L2 length)}

\[ \|u\| = \sqrt{\sum_{i=1}^{N} u_i^2} \]

\[ \text{doc} x \cdot \text{doc} y = \sqrt{2^2 + 4^2} = 4.47 \]
Vector L2 (length) normalization

Normalization of \( \mathbf{u} \) is a vector \( \mathbf{\hat{u}} \) obtained by dividing each element of \( \mathbf{u} \) by \( ||\mathbf{u}|| \):

\[
||\mathbf{u}|| = \sqrt{\sum_{i=1}^{N} u_i^2}
\]

\[
\mathbf{\hat{u}} = \frac{1}{||\mathbf{u}||} \mathbf{u}
\]

\[
\text{doc x doc y} \quad \leq \quad \sqrt{2^2 + 4^2} = 4.47
\]

\[
\begin{align*}
A & : 0.45 \quad 0.89 \\
B & : 0.55 \quad 0.93 \\
c & : 0.81 \quad 0.58
\end{align*}
\]

\[
\mathbf{A} = \left[ \frac{2}{4.47}, \frac{u}{4.47} \right]
\]

Cosine distance

\[
1 - \frac{\sum_{i=1}^{N} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \cdot \sqrt{\sum_{i=1}^{N} v_i^2}}
\]

What information do we lose when we discard freq?

Distributional inclusion hypothesis

A is a B

dire wolf is a animal

How do I make a nice graph \( \text{dim} \gg 2 \)?
Matrix type

Vector comparisons
- cosine
- Manhattan distance
- KL-divergence
- JS distance
- Dice coefficient
- Jaccard

Reweighting
- length norm.
- probabilities
- TF-IDF
- PMI
- positive PMI

Dim. Reduction
- LSA
- GloVe
- PCA
- word2vec
- CBOW
- skip-gram

How many dimensions should we use?
- 10?
- 50?
- 100?

How to pick the right combination?
- Trial and error.

Evaluating word vectors

2 kinds of evaluation
- extrinsic evaluation = task-based
- intrinsic

Psycholinguistic experiment

<table>
<thead>
<tr>
<th>Word Pair</th>
<th>Score</th>
<th>Mean 10 Judges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love, sex</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>tiger, cat</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>tiger, tiger</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>fertility, egg</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>stock, eggs</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>professor, cucumber</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

WordSim353

GOLD STANDARD
Compute correlation

- Spearman's rank correlation coefficient
  - Kendall's tau
  
  \[ \tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\frac{N(N-1)}{2}} \]

Human ordering:

- Sex, love
- Prof, cucumber

System ordering predicts:

- \( \Rightarrow \) discordant pair
- \( \Rightarrow \) concordant pair

range \( \rightarrow \) to \( \rightarrow \)

wordvectors.org

Community based eval

leaderboard

visualizations + t-SNE

Does similarity == meaning/semantics?

- \( \text{dog}(x) \Rightarrow \text{animal}(x) \)
- \( \text{dog}(x) \Rightarrow \neg \text{gorilla}(x) \)

Word vectors fail to capture entailment implications
Does similarity == meaning/semantics?

\[
\begin{align*}
\text{dog}(x) & \Rightarrow \text{animal}(x) \\
\text{dog}(x) & \Rightarrow \neg \text{gorilla}(x)
\end{align*}
\]

Word vectors fail to capture logical implications.

\[
\text{sim(boys, girls)} \quad \text{high sim}
\]

\[
\begin{align*}
\text{cats} & \quad \text{dogs} \\
\text{France} & \quad \text{Germany} \\
\text{rise} & \quad \text{fall}
\end{align*}
\]

Next time:

How Word2Vec can be used to solve analogies!