A Simple Semantic Space

Stuart B. Opotowsky was named vice president for this company with interests in insurance, tobacco, hotels and broadcasting.

Outline

1 Introduction
   - Semantic Space Models
   - Logic-based View

2 Composition Models

3 Evaluation
   - Phrase Similarity Task
   - Paraphrase Detection

A Simple Semantic Space

Stuart B. Opotowsky was named vice president for this company with interests in insurance, tobacco, hotels and broadcasting.

Select 2,000 most common content words as contexts.
Stuart B. Opotowsky was named vice president for this company with interests in insurance, tobacco, hotels and broadcasting.

- Select 2,000 most common content words as contexts.
- Five word context window each side of the target word.

<table>
<thead>
<tr>
<th>company</th>
<th>vice</th>
<th>president</th>
<th>tax</th>
<th>interests</th>
<th>insurance</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
<td>103</td>
<td>19</td>
<td>55</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Select 2,000 most common content words as contexts.
- Five word context window each side of the target word.

<table>
<thead>
<tr>
<th>company</th>
<th>vice</th>
<th>president</th>
<th>tax</th>
<th>interests</th>
<th>insurance</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.06</td>
<td>0.26</td>
<td>0.05</td>
<td>0.14</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

- Convert counts to probabilities: $p(c|w)$.
- Divide through by probabilities of each context word: $p(c|w) / p(c)$.

Cosine similarity: $\text{sim}(w_1, w_2) = w_1 \cdot w_2 / |w_1| |w_2|$.
A Simple Semantic Space

<table>
<thead>
<tr>
<th></th>
<th>vice</th>
<th>president</th>
<th>tax</th>
<th>interests</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>company</td>
<td>1.52</td>
<td>2.32</td>
<td>1.14</td>
<td>1.06</td>
<td>...</td>
</tr>
</tbody>
</table>

- Select 2,000 most common content words as contexts.
- Five word context window each side of the target word.
- Convert counts to probabilities: $p(c|w)$.
- Divide through by probabilities of each context word: $\frac{p(c|w)}{p(c)}$.

Cosine similarity: $\text{sim}(w_1, w_2) = \frac{w_1 \cdot w_2}{||w_1|| ||w_2||}$.

Distributional Semantics

Words are represented through their relations to other words.

Topic Models

**Key Idea:** documents are mixtures of topics, topics are probability distributions over words (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004).

Topic models are **generative** and **structured**. For a new document:

- Choose a distribution over topics
- Choose a topic at random according to distribution
- Draw a word from that topic

Statistical techniques used to invert the process: infer set of topics that were responsible for generating a collection of documents.
2. Generative Models

**Probabilistic Generative Process**

- **DOC1**: money\(^1\) bank\(^1\) loan\(^1\)
  - money\(^1\)
  - bank\(^1\)
  - loan\(^1\)
- **DOC2**: money\(^1\) bank\(^1\)
  - river\(^2\) loan\(^2\) stream\(^2\)
  - bank\(^1\) money\(^1\)
- **DOC3**: river\(^2\) bank\(^2\)
  - stream\(^2\) bank\(^2\) river\(^2\) stream\(^2\) bank\(^2\)

**TOPIC 1**
- *money, bank, loan*

**TOPIC 2**
- *river, stream, bank*

---

**Statistical Inference**

- **DOC1**: money\(^2\) bank\(^2\) loan\(^2\)
  - bank\(^2\)
  - money\(^2\)
  - loan\(^2\)
- **DOC2**: money\(^2\) bank\(^2\)
  - river\(^2\) loan\(^2\) stream\(^2\)
  - bank\(^2\) money\(^2\)
- **DOC3**: river\(^2\) bank\(^2\)
  - stream\(^2\) bank\(^2\) river\(^2\) stream\(^2\) bank\(^2\)

**TOPIC 1**
- *money, bank, loan*

**TOPIC 2**
- *river, stream, bank*

---

**Meaning Representation**

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic n</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0.39</td>
<td>0.02</td>
</tr>
<tr>
<td>difficulty</td>
<td>0.03</td>
<td>0.44</td>
</tr>
<tr>
<td>produce</td>
<td>0.06</td>
<td>0.17</td>
</tr>
</tbody>
</table>

- Topics correspond to coarse-grained sense distinctions
- Cosine similarity can be used (probabilistic alternatives)

---

**Semantic Space Models**

Semantic space models are extremely popular across disciplines!

- Semantic Priming (Lund and Burgess, 1996)
- Text comprehension (Landauer and Dumais, 1997)
- Word association (McDonald, 2000)
- Information Retrieval (Salton et al., 1975)
- Thesaurus extraction (Grefenstette, 1994)
- Word Sense disambiguation (Schütze, 1998)
- Text Segmentation (Hirst, 1997)
- **Automatic, language independent**
Semantic Space Models

Semantic space models are extremely popular across disciplines!

- Semantic Priming (Lund and Burgess, 1996)
- Text comprehension (Landauer and Dumais, 1997)
- Word association (McDonald, 2000)
- Information Retrieval (Salton et al., 1975)
- Thesaurus extraction (Grefenstette, 1994)
- Word Sense disambiguation (Schütze, 1998)
- Text Segmentation (Hirst, 1997)
- **Automatic, language independent**

_Catch:_ representation of the meaning of _single words_. What about _phrases_ or _sentences_?

---

Quick Fix

It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.

That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

- Vector averaging: \( p = \frac{1}{2} (u + v) \) (Foltz et al., 1998; Landauer et al., 1997); _syntax insensitive_

---

Quick Fix

It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.

That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

- Vector averaging: \( p = \frac{1}{2} (u + v) \) (Foltz et al., 1998; Landauer et al., 1997); _syntax insensitive_
- Add a neighbor to the sum: \( p = u + v + n \) (Kintsch, 2001); _meaning of predicate depends on its argument_
Meaning of whole is the meaning of its parts (Frege, 1957).

\[
\lambda u.\lambda v. \exists x (u \circ x \land v \circ x) \quad \lambda y. \text{HORSE}(y) \quad \lambda z. \text{RUN}(z)
\]

- Logic accounts for sentential meaning (Montague, 1974).
- Differences are qualitative rather than quantitative.
- Cannot express degrees of similarity.
Compositionality

Partee (1995): the meaning of the whole is a function of the meaning of the parts and of the way they are **syntactically** combined.

Lakoff (1977): the meaning of the whole is a **greater** than the meaning of the parts.

Frege (1884): never ask the meaning of a word in **isolation** but only in the **context** of a statement.

Pinker (1994): composition of simple elements must allow the construction of **novel meanings** which go beyond those of the individual elements.
- Tensor products: $p = u \otimes v$ (Smolensky, 1990); **dimensionality**
- Circular convolution: $p = u \odot v$ (Plate, 1991); **components are randomly distributed**
- Spatter codes: take the XOR of two vectors (Kanerva, 1998); **components are random bits**

**A Framework for Semantic Composition**

$$p = f(u, v, R, K)$$
A Framework for Semantic Composition

\[ p = f(u, v, R, K) \]

- composition of \( u, v \)
- syntactic relationship

**Assumptions:**
1. Eliminate background knowledge \( K \)
2. Vary syntactic relationship \( R \)
3. \( p \) is in the same space as \( u \) and \( v \)
4. \( f() \) is a linear function of Cartesian product (additive)
5. \( f() \) is a linear function of tensor product (multiplicative)
A Framework for Semantic Composition

\[ p = f(u, v, OBJ) \]

composition of \( u, v \)
syntactic relationship

Assumptions:
1. eliminate background knowledge \( K \)
2. vary syntactic relationship \( R \)
3. \( p \) is in same space as \( u \) and \( v \)
4. \( f() \) is a linear function of Cartesian product (additive)
5. \( f() \) is a linear function of tensor product (multiplicative)
Additive Models

\[ p = Au + Bv \]

Instances

\[ p = u + v \]
\[ p = u + v + \sum_{i} n_i \]
\[ p = \alpha u + \beta v \]
\[ p = v \]

\[
\begin{array}{l}
\text{music} & \text{solution} & \text{economy} & \text{craft} & \text{create} \\
\hline
\text{practical} & 0 & 6 & 2 & 10 & 4 \\
\text{difficulty} & 1 & 8 & 4 & 4 & 0 \\
\text{problem} & 2 & 15 & 7 & 9 & 1 \\
\end{array}
\]

Practical + Difficulty = [1 14 6 14 4]

Practical + Difficulty + Problem = [3 29 13 23 5]

0.4 · Practical + 0.6 · Difficulty = [0.6 5.6 3.2 6.4 1.6]
### Additive Models

- \( p = Au + Bv \)
- \( p = u + v \)
- \( p = u + v + \sum_i n_i \)
- \( p = \alpha u + \beta v \)
- \( p = v \)

### Instances

<table>
<thead>
<tr>
<th>practical</th>
<th>difficulty</th>
<th>problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>difficulty</td>
<td>problem</td>
</tr>
</tbody>
</table>

### Evaluation

- \( p = Au + Bv \)
- \( p = u + v \)

### Multiplicative Models

- \( p = Cuv \)
- \( p = u \odot v \)
- \( p = u \odot v \)
- \( p = u \odot v \)

### Instances

<table>
<thead>
<tr>
<th>practical</th>
<th>difficulty</th>
<th>problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>difficulty</td>
<td>problem</td>
</tr>
</tbody>
</table>

### Equation Examples

- \( p = \alpha u + \beta v \)
- \( p = v \)
- \( p = u \odot v \)
- \( p = u \odot v \)

### Model Solutions

- Practical + difficulty = [1 14 6 14 4]
- Practical + difficulty + problem = [3 29 13 23 5]
- \( 0.4 \cdot \text{practical} + 0.6 \cdot \text{difficulty} = [0.6 5.6 3.2 6.4 1.6] \)
- Practical + difficulty = [1 8 4 4 0]
- Difficult = [1 8 4 4 0]

### Multiplicative Model Solutions

- Practical \( \odot \) difficulty = [0 48 8 40 0]
- Practical \( \odot \) difficulty = [0 48 8 40 0]
Models (Mitchell and Lapata, 2010)

Multiplicative Models

\[ p = C_{uv} \]

**Instances**

\[ p = u \odot v \]
\[ \rho_i = u_i \cdot v_i \]
\[ p = u \otimes v \]
\[ \rho_{i,j} = u_i \cdot v_{i-j} \]

**practical \odot difficulty**

<table>
<thead>
<tr>
<th>practical</th>
<th>0</th>
<th>6</th>
<th>2</th>
<th>10</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

**Evaluation**

**Dilation Models**

\[ p = C_{uv} = Uv \]
\[ p = (\lambda - 1)(u \cdot v)u + (u \cdot u)v \]

**Dilation Models**

\[ p = C_{uv} = Uv \]
\[ p = (\lambda - 1)(u \cdot v)u + (u \cdot u)v \]
**Dilation Models**

\[ p = C_{uv} = Uv \]

\[ U_{ij} = 0, U_{ii} = u_i \]

\[ x = \frac{u \cdot v}{u \cdot u} u \quad y = v - x = v - \frac{u \cdot v}{u \cdot u} u \]

\[ v' = \lambda x + y = (\lambda - 1) \frac{u \cdot v}{u \cdot u} u + v \]

\[ p = (\lambda - 1)(u \cdot v)u + (u \cdot u)v \]

---

**Phrase Similarity Task** (Kintsch, 2002)

- Elicit similarity judgments for adjective-noun, noun-noun, verb-object combinations.
- Phrase pairs from three bands: High, Medium, Low.
- Compute vectors for phrases, measure their similarity.
- Correlate model similarities with human ratings.
Elicit similarity judgments for adjective-noun, noun-noun, verb-object combinations.
Phrase pairs from three bands: High, Medium, Low.
Compute vectors for phrases, measure their similarity.
Correlate model similarities with human ratings.

High
- old person
- kitchen door
- produce effect

Medium
- elderly lady
- right hand
- small house

Low
- bedroom window
- office worker
- housing department
- produce effect
- achieve result
- consider matter
- start work

Results for verb-obj (Mitchell and Lapata, 2010)
Results for noun-noun (Mitchell and Lapata, 2010)

Summary

- Multiplicative and dilation models best for simple space
- Dilation and additive models best for LDA model
- Circular convolution is worst performing model
- Different composition functions appropriate for different representations (additive vs. multiplicative)

What are composition models good for?
- modeling brain activity (Chang et al., 2009)
- language modeling (Mitchell and Lapata, 2009)
- modeling eye tracking data (Mitchell et al., 2010)
- paraphrase detection (Blacoe and Lapata, 2011)

Paraphrase Detection

Given: A pair of sentences $S_1 = (w_1 \ldots w_m)$ and $S_2 = (w_1 \ldots w_n)$

Task: Classify whether $S_1$ and $S_2$ are paraphrases or not

Amrozi accused his brother, whom he called “the witness”, of deliberately distorting his evidence.

Referring to him as only “the witness”, Amrozi accused his brother of deliberately distorting his evidence.
Paraphrase Detection

- Microsoft Research Paraphrase Corpus (Dolan et al., 2004).
- Features: sentence vectors concatenated, subtracted, encoding of words in sentence, sentence vector similarity, unigram overlap, sentence lengths

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.5</td>
<td>79.9</td>
</tr>
<tr>
<td>Mihalcea et al. (2006)</td>
<td>70.3</td>
<td>81.3</td>
</tr>
<tr>
<td>Rus et al. (2008)</td>
<td>70.6</td>
<td>80.5</td>
</tr>
<tr>
<td>Qiu et al. (2006)</td>
<td>72.0</td>
<td>81.6</td>
</tr>
<tr>
<td>Islam et al. (2007)</td>
<td>72.6</td>
<td>81.3</td>
</tr>
<tr>
<td><strong>Mitchell and Lapata (2010)</strong></td>
<td><strong>73.0</strong></td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>Fernando et al. (2008)</td>
<td>74.1</td>
<td>82.4</td>
</tr>
<tr>
<td>Wan et al. (2006)</td>
<td>75.6</td>
<td>83.0</td>
</tr>
<tr>
<td>Socher et al. (76.4)</td>
<td>76.4</td>
<td>83.6</td>
</tr>
</tbody>
</table>

LDA Topics

![LDA Topics Graph](image-url)