Vector-based Models of Semantic Composition

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk
1 Introduction
   - Semantic Space Models
   - Logic-based View
   - Connectionism

2 Composition Models

3 Evaluation
   - Phrase Similarity Task
   - Paraphrase Detection

4 Conclusions
You shall know a word by the company it keeps (Firth, 1957)
A word’s context provides information about its meaning
Words are similar if they share similar linguistic contexts
Stuart B. Opotowsky was named vice president for this **company** with interests in insurance, tobacco, hotels and broadcasting.
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.
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.
- Five word context window each side of the target word.
A Simple Semantic Space

- Select 2,000 most common content words as contexts.
- Five word context window each side of the target word.
### 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>25</td>
<td>103</td>
<td>19</td>
<td>55</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.
A Simple Semantic Space

Select 2,000 most common content words as contexts.

Five word context window each side of the target word.

Convert counts to probabilities: $p(c|w)$.
A Simple Semantic Space

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: $\frac{p(c|w)}{p(c)}$.

<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>
## 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: $\text{sim}(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{|\mathbf{w}_1||\mathbf{w}_2|}$. 
Distributional Semantics

Words are represented through their relations to other words.
**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:

1. Choose a distribution over topics
2. Choose a topic at random according to distribution
3. 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.
Probabilistic Generative Process

TOPIC 1

money
loan
bank

1.0

DOC1: money¹ bank¹ loan¹
       bank¹ money¹ money¹
       bank¹ loan¹

.5

DOC2: money¹ bank¹
       bank² river² loan¹ stream²
       bank¹ money¹

.5

DOC3: river² bank²
       stream² bank² river² river²
       stream² bank²

TOPIC 2

river
bank
stream
Statistical Inference

Figure 2.

TOPIC 1


TOPIC 2


### Meaning Representation

<table>
<thead>
<tr>
<th>Word</th>
<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>
<td>...</td>
</tr>
<tr>
<td>difficulty</td>
<td>0.03</td>
<td>0.44</td>
<td>...</td>
</tr>
<tr>
<td>produce</td>
<td>0.06</td>
<td>0.17</td>
<td>...</td>
</tr>
</tbody>
</table>

Topics are the dimensions of the space (500, 1000)

- Vector components: probability of word given topic
- Topics correspond to coarse-grained sense distinctions
- Cosine similarity can be used (probabilistic alternatives)
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 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.
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: $\mathbf{p} = \frac{1}{2}(\mathbf{u} + \mathbf{v})$ (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: \( \mathbf{p} = \frac{1}{2}(\mathbf{u} + \mathbf{v}) \) (Foltz et al., 1998; Landauer et al., 1997); syntax insensitive

- Add a neighbor to the sum: \( \mathbf{p} = \mathbf{u} + \mathbf{v} + \mathbf{n} \) (Kintsch, 2001); meaning of predicate depends on its argument
Logic-based View

Meaning of whole is function of meaning of its parts (Frege, 1957).
Meaning of whole is function of meaning of its parts (Frege, 1957).

\[
\lambda u. \lambda v. \exists x (u@x \land v@x) \quad \lambda y. \text{HORSE}(y) \quad \lambda z. \text{RUN}(z)
\]
Meaning of whole is function of meaning of its parts (Frege, 1957).

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

\[ \exists x (\text{HORSE}(x) \land \text{RUN}(x)) \]
Logic-based View

Meaning of whole is function of meaning of its parts (Frege, 1957).

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

\[
\exists x (\text{HORSE}(x) \land \text{RUN}(x))
\]

- Logic can account for sentential meaning (Montague, 1974).
- Differences in meaning 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.
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.
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.
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); \textbf{dimensionality}
Connectionism

Tensor products: $p = u \otimes v$ (Smolensky, 1990); \textbf{dimensionality}

Circular convolution: $p = u \odot v$ (Plate, 1991); \textbf{components are randomly distributed}
Tensor products: $p = u \otimes v$ (Smolensky, 1990); dimensionality
Circular convolution: $p = u \ast v$ (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

Composition of $u, v$

$$p = f(u, v, R, K)$$
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 model)
5. \( f() \) is a linear function of tensor product (multiplicative model)
A Framework for Semantic Composition

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

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

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 model)
5. \( f() \) is a linear function of tensor product (multiplicative model)
Composition Models

A Framework for Semantic Composition

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

composition of \( u, v \)

syntactic relationship

Assumptions:

1. eliminate background knowledge \( K \)
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 \( R \)
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 \( v \)
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 model)
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 model)
5. \( f() \) is a linear function of tensor product (multiplicative model)
Composition Models

Models (Mitchell and Lapata, 2010)

Additive Models

\[ p = Au + Bv \]

Instances

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

\[ p = Au + Bv \]

**Instances**

\[ p = u + v \]

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

\[ p = \alpha u + \beta v \]

\[ p = v \]

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

**practical + difficulty** = \[1 \ 14 \ 6 \ 14 \ 4\]
Additive Models

\[ p = Au + Bv \]

Instances

\[ p = u + v \]

\[ p = u + v + \sum n_i \]

\[ p = \alpha u + \beta v \]

\[ p = v \]

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

practical + difficulty = [1 14 6 14 4]

practical + difficulty + problem = [3 29 13 23 5]
Composition Models

Models (Mitchell and Lapata, 2010)

Additive Models

\[ p = Au + Bv \]

Instances

\[ p = u + v \]

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

\[ p = \alpha u + \beta v \]

\[ p = v \]

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\text{practical} + \text{difficulty} = [1 \ 14 \ 6 \ 14 \ 4]
\]

\[
\text{practical} + \text{difficulty} + \text{problem} = [3 \ 29 \ 13 \ 23 \ 5]
\]

\[
0.4 \cdot \text{practical} + 0.6 \cdot \text{difficulty} = [0.6 \ 5.6 \ 3.2 \ 6.4 \ 1.6]
\]
### Composition Models

#### Models (Mitchell and Lapata, 2010)

- **Additive Models**
  
  \[ p = Au + Bv \]

- **Instances**
  
  \[ p = u + v \]

  \[ p = u + v + \sum_i n_i \]

  \[ p = \alpha u + \beta v \]

  \[ p = v \]

<table>
<thead>
<tr>
<th>Music</th>
<th>Solution</th>
<th>Economy</th>
<th>Craft</th>
<th>Create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

**Calculations**

- \( \text{practical} + \text{difficulty} = [1 \ 14 \ 6 \ 14 \ 4] \)

- \( \text{practical} + \text{difficulty} + \text{problem} = [3 \ 29 \ 13 \ 23 \ 5] \)

- \( 0.4 \cdot \text{practical} + 0.6 \cdot \text{difficulty} = [0.6 \ 5.6 \ 3.2 \ 6.4 \ 1.6] \)

- \( \text{difficulty} = [1 \ 8 \ 4 \ 4 \ 0] \)
**Multiplicative Models**

\[ p = Cuv \]

**Instances**

\[ p = u \odot v \]
\[ p_i = u_i v_i \]

\[ p = u \otimes v \]
\[ p_{i,j} = u_i \cdot v_j \]

\[ p = u \oslash v \]
\[ p_i = \sum_j u_j \cdot v_{i-j} \]
**Multiplicative Models**

\[ p = C_{uv} \]

**Instances**

\[ p = u \odot v \]
\[ p_i = u_i v_i \]

\[ p = u \otimes v \]
\[ p_{i,j} = u_i \cdot v_j \]

\[ p = u \oslash v \]
\[ p_i = \sum_j u_j \cdot v_{i-j} \]

<table>
<thead>
<tr>
<th>practical</th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>difficulty</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>practical</td>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

**practical \odot difficulty** = \[ \begin{bmatrix} 0 & 48 & 8 & 40 & 0 \end{bmatrix} \]
### Multiplicative Models

\[ p = Cuv \]

#### Instances

\[ p = u \odot v \]
\[ p_i = u_i v_i \]

\[ p = u \otimes v \]
\[ p_{i,j} = u_i \cdot v_j \]

\[ p = u \oslash v \]
\[ p_i = \sum_j u_j \cdot v_{i-j} \]

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{practical} \odot \text{difficulty} = [0 \ 48 \ 8 \ 40 \ 0] \]

\[ \text{practical} \otimes \text{difficulty} = \]
\[
\begin{array}{ccccc}
0 & 0 & 0 & 0 & 0 \\
6 & 48 & 24 & 24 & 0 \\
2 & 16 & 8 & 8 & 0 \\
10 & 80 & 40 & 40 & 0 \\
4 & 32 & 16 & 16 & 0 \\
\end{array}
\]


**Models** (Mitchell and Lapata, 2010)

**Multiplicative Models**

\[ p = Cuv \]

**Instances**

\[ p = u \odot v \]

\[ p_i = u_i v_i \]

\[ p = u \otimes v \]

\[ p_{i,j} = u_i \cdot v_j \]

\[ p = u \oslash v \]

\[ p_i = \sum_j u_j \cdot v_{i-j} \]

<table>
<thead>
<tr>
<th>( \text{music} )</th>
<th>( \text{solution} )</th>
<th>( \text{economy} )</th>
<th>( \text{craft} )</th>
<th>( \text{create} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

\[ \text{practical} \odot \text{difficulty} = [0 \ 48 \ 8 \ 40 \ 0] \]

\[ \text{practical} \otimes \text{difficulty} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 6 & 48 & 24 & 24 & 0 \\ 2 & 16 & 8 & 8 & 0 \\ 10 & 80 & 40 & 40 & 0 \\ 4 & 32 & 16 & 16 & 0 \end{bmatrix} \]

\[ \text{practical} \oslash \text{difficulty} = [116 \ 50 \ 66 \ 62 \ 80] \]
Composition Models

Models (Mitchell and Lapata, 2010)

Dilation Models

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

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

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

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

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

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

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

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

\[ p = C uv = U v \]
\[ U_{ij} = 0, \ U_{ii} = u_i \]

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

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

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

\[ p = Cuv = Uv \]
\[ U_{ij} = 0, \quad U_{ii} = u_i \]

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

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

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

\[ p = Cu v = Uv \]
\[ U_{ij} = 0, U_{ii} = u_i \]

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

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

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

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>old person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kitchen door</td>
<td></td>
<td></td>
</tr>
<tr>
<td>produce effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>produce effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>achieve result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>produce effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>office worker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>housing department</td>
<td></td>
<td></td>
</tr>
<tr>
<td>produce effect</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mirella Lapata
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.

<table>
<thead>
<tr>
<th>Old person</th>
<th>Elderly lady</th>
<th>Right hand</th>
<th>Small house</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen door</td>
<td>Produce effect</td>
<td>Office worker</td>
<td>Housing department</td>
</tr>
<tr>
<td>Bedroom window</td>
<td></td>
<td>Start work</td>
<td>Achieve result</td>
</tr>
<tr>
<td>Produce effect</td>
<td></td>
<td>Consider matter</td>
<td>Effect</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>old person</td>
<td>elderly lady</td>
<td>small house</td>
</tr>
<tr>
<td>kitchen door</td>
<td>bedroom window</td>
<td>housing department</td>
</tr>
<tr>
<td>produce effect</td>
<td>right hand</td>
<td></td>
</tr>
<tr>
<td>effect</td>
<td>office worker</td>
<td></td>
</tr>
</tbody>
</table>
## 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.

<table>
<thead>
<tr>
<th></th>
<th><strong>High</strong></th>
<th><strong>Medium</strong></th>
<th><strong>Low</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>old person</td>
<td>elderly lady</td>
<td>right hand</td>
<td>small house</td>
</tr>
<tr>
<td>kitchen door</td>
<td>bedroom window</td>
<td>office worker</td>
<td>housing department</td>
</tr>
<tr>
<td>produce effect</td>
<td>achieve result</td>
<td>consider matter</td>
<td>start work</td>
</tr>
</tbody>
</table>
Results for verb-obj (Mitchell and Lapata, 2010)

![Graph showing Spearman's rho values for various semantic models](chart.png)
Results for verb-obj (Mitchell and Lapata, 2010)
Results for verb-obj (Mitchell and Lapata, 2010)

![Graph showing comparison between Human Upper Bound, Simple Semantic Space, and LDA Space for various operations like Add, Kintsch, WAdd, Multiply, TensorP, CConv, Dilation, and HeadO. The graph plots Spearman’s rho values.]
Results for adj-noun (Mitchell and Lapata, 2010)

The diagram shows the Spearman's rho values for various methods in a phrase similarity task. The methods include:

- **Add**
- **Kintsch**
- **Multiply**
- **TensorP**
- **CConv**
- **Wadd**
- **Dilation**
- **HeadO**

The metrics compared are:

- **Human Upper Bound**
- **Simple Semantic Space**
- **LDA**

The graph illustrates the performance of these methods, with 'Add' showing the highest similarity values among the human upper bound and simple semantic space methods.
Results for noun-noun (Mitchell and Lapata, 2010)

- Add
- Kintsch
- Multiply
- Tensor
- Conv
- Wadd
- Dilation
- HeadO

Spearman’s rho

- Human Upper Bound
- Simple Semantic Space
- LDA

Mirella Lapata
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)

- 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)
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

- The lies and deceptions from Saddam have been well documented over 12 years.
- It has been well documented over 12 years of lies and deception from Saddam.
Paraphrase Detection

- Microsoft Research Paraphrase Corpus (Dolan et al., 2004).
- Features: **sentence vectors concatenated** (con), **subtracted** (sub), encoding of words in sentence (enc), **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>73.0</td>
<td>83.3</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>
Conclusions

- University of Edinburgh
  (http://homepages.inf.ed.ac.uk/s1066731/)

- Stanford University, Richard Socher
  (http://www.socher.org/)

- University of Trento, COMPOSES
  (http://clic.cimec.unitn.it/composes/)

- Universities of Oxford and Cambridge, DiscoCat

- Carnegie Mellon University (Alona Fyshe, Brian Murphy)

- Workshop at ACL-2013: Continuous Vector Space Models and their Compositionality
LDA Topics

The graph shows the perplexity of additive and multiplicative LDA models as a function of the number of topics. The perplexity decreases significantly with the addition of topics for the additive model, while the multiplicative model shows a more gradual increase.

- **Additive**
  - Initially decreases sharply with the addition of topics.
  - Continues to decrease but at a slower rate.

- **Multiplicative**
  - Increases gradually with the addition of topics.
  - Stabilizes after a certain number of topics.

The graph indicates that the additive model converges more quickly to a lower perplexity, making it more efficient for topic modeling compared to the multiplicative model.