Lecture 4: Models of Semantic Composition

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

January 27, 2016

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.

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

	vice	president	interests	insurance	...
company	1	1	1	1	...

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

A Simple Semantic Space

	vice	president	tax	interests	...
company	25	103	19	55	...

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

A Simple Semantic Space

	vice	president	tax	interests	...
company	0.06	0.26	0.05	0.14	...

- 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

	vice	president	tax	interests	...
company	1.52	2.32	1.14	1.06	...

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

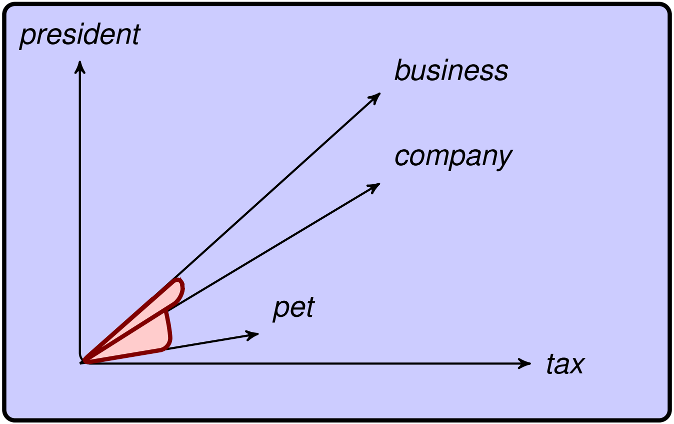
A Simple Semantic Space

	vice	president	tax	interests	...
company	1.52	2.32	1.14	1.06	...

- 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: $sim(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{|\mathbf{w}_1||\mathbf{w}_2|}$.

Distributional Semantics

Words are represented through their relations to other words.



Topic Models

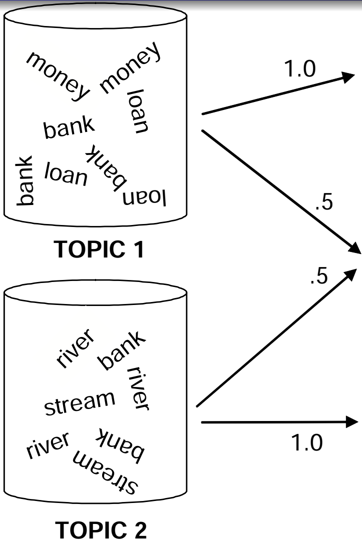
Key Idea: documents are mixtures of topics, topics are probability distributions over words (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004).

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

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

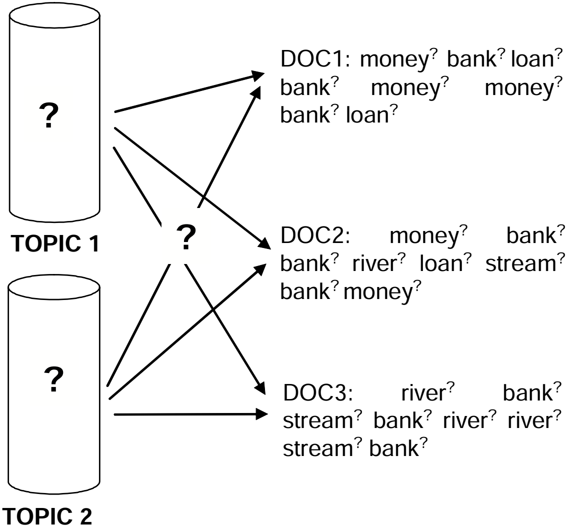


DOC1: money^1 bank^1 loan^1 bank^1 money^1 money^1 bank^1 loan^1

DOC2: money^1 bank^1 bank^2 river^2 loan^1 stream^2 bank^1 money^1

DOC3: river^2 bank^2 stream^2 bank^2 river^2 river^2 stream^2 bank^2

Statistical Inference



Meaning Representation

	Topic 1	Topic 2	Topic n
practical	0.39	0.02	...
difficulty	0.03	0.44	...
produce	0.06	0.17	...

- 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

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.

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 the meaning of its parts (Frege, 1957).

Logic-based View

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

$$\begin{array}{ccc} \textbf{a} & \textbf{horse} & \textbf{ran} \\ \downarrow & \downarrow & \downarrow \\ \lambda u.\lambda v.\exists x(u@x \wedge v@x) & \lambda y.HORSE(y) & \lambda z.RUN(z) \end{array}$$

Logic-based View

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

$$
\begin{tikzcd}
\textbf{a} \arrow[d] & \textbf{horse} \arrow[d] & \textbf{ran} \arrow[d] \\
\lambda u.\lambda v.\exists x(u@x \wedge v@x) \arrow[dr] & \lambda y.HORSE(y) \arrow[d] & \lambda z.RUN(z) \arrow[dl] \\
& \exists x(HORSE(x) \wedge RUN(x)) &
\end{tikzcd}
$$

Logic-based View

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

a | **horse** | **ran**

$\lambda u.\lambda v.\exists x(u@x \wedge v@x)$ | $\lambda y.HORSE(y)$ | $\lambda z.RUN(z)$

$$\exists x(HORSE(x) \wedge RUN(x))$$

- 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.

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

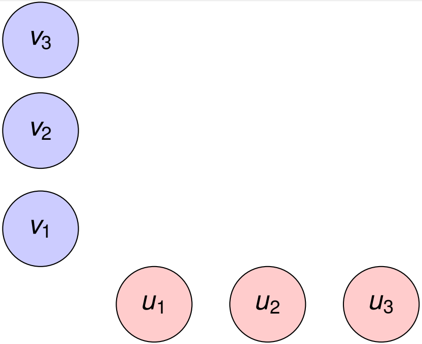
Partee (1995): the meaning of the whole is a function of the meaning of the parts and of the way they are **syntactically** combined.

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.

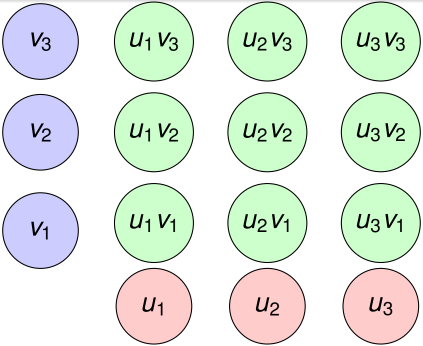
Pinker (1994): composition of simple elements must allow the construction of **novel meanings** which go beyond those of the individual elements.

Connectionism



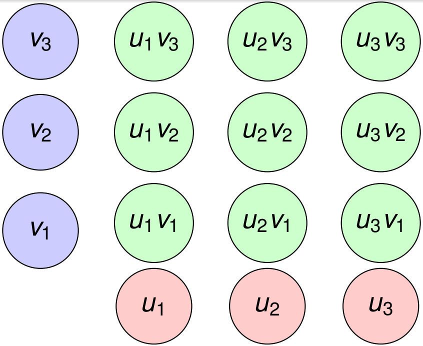
- Tensor products: $\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$ (Smolensky, 1990); **dimensionality**

Connectionism



- Tensor products: $\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$ (Smolensky, 1990); **dimensionality**
- Circular convolution: $\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$ (Plate, 1991); **components are randomly distributed**

Connectionism



- Tensor products: $\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$ (Smolensky, 1990); **dimensionality**
- Circular convolution: $\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$ (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

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K)}$$

A Framework for Semantic Composition

$$\boxed{\color{red}{\mathbf{p}} = f(\mathbf{u}, \mathbf{v}, R, K)}$$

composition of $\mathbf{u}, \mathbf{v}$

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K)}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K)}$$

composition of $\mathbf{u}, \mathbf{v}$

background knowledge

syntactic relationship

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R)}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

Assumptions:

1. eliminate background knowledge K

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, \textcolor{red}{OBJ})}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

Assumptions:

1. eliminate background knowledge K
2. vary syntactic relationship R

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, \mathit{OBJ})}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

Assumptions:

1. eliminate background knowledge K
2. vary syntactic relationship R
3. $\mathbf{p}$ is in same space as $\mathbf{u}$ and $\mathbf{v}$

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, OBJ)}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

Assumptions:

1. eliminate background knowledge K
2. vary syntactic relationship R
3. $\mathbf{p}$ is in same space as $\mathbf{u}$ and $\mathbf{v}$
4. $f()$ is a linear function of Cartesian product (**additive**)

A Framework for Semantic Composition

$$\boxed{\mathbf{p} = f(\mathbf{u}, \mathbf{v}, OBJ)}$$

composition of $\mathbf{u}, \mathbf{v}$

syntactic relationship

Assumptions:

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

Models (Mitchell and Lapata, 2010)

Additive Models

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$$

Instances

$$\mathbf{p} = \mathbf{u} + \mathbf{v}$$

$$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i$$

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$$

$$\mathbf{p} = \mathbf{v}$$

Models (Mitchell and Lapata, 2010)

Additive Models

$\mathbf{p} = \mathbf{Au} + \mathbf{Bv}$

Instances

$\mathbf{p} = \mathbf{u} + \mathbf{v}$

$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i$

$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$

$\mathbf{p} = \mathbf{v}$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

practical $+$ **difficulty** $=$ [1 14 6 14 4]

Models (Mitchell and Lapata, 2010)

Additive Models

$\mathbf{p} = \mathbf{Au} + \mathbf{Bv}$

Instances

$\mathbf{p} = \mathbf{u} + \mathbf{v}$

$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i$

$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$

$\mathbf{p} = \mathbf{v}$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

practical + **difficulty** = [1 14 6 14 4]

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

Models (Mitchell and Lapata, 2010)

Additive Models

$\mathbf{p} = \mathbf{Au} + \mathbf{Bv}$

Instances

$\mathbf{p} = \mathbf{u} + \mathbf{v}$

$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i$

$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$

$\mathbf{p} = \mathbf{v}$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

practical + **difficulty** = [1 14 6 14 4]

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

$0.4 \cdot$ **practical** $+ 0.6 \cdot$ **difficulty** = [0.6 5.6 3.2 6.4 1.6]

Models (Mitchell and Lapata, 2010)

Additive Models

$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$

Instances

$\mathbf{p} = \mathbf{u} + \mathbf{v}$

$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i$

$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$

$\mathbf{p} = \mathbf{v}$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

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

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

$0.4 \cdot \textbf{practical} + 0.6 \cdot \textbf{difficulty} = [0.6\ 5.6\ 3.2\ 6.4\ 1.6]$

$\textbf{difficulty} = [1\ 8\ 4\ 4\ 0]$

Models (Mitchell and Lapata, 2010)

Multiplicative Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v}$$

Instances

$$\mathbf{p} = \mathbf{u} \odot \mathbf{v}$$
$$p_i = u_i v_i$$

$$\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$$
$$p_{i,j} = u_i \cdot v_j$$

$$\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$$
$$p_i = \sum_j u_j \cdot v_{i-j}$$

Models (Mitchell and Lapata, 2010)

Multiplicative Models

$$\mathbf{p} = \mathbf{Cuv}$$

Instances

$$\mathbf{p} = \mathbf{u} \odot \mathbf{v}$$
$$p_i = u_i v_i$$

$$\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$$
$$p_{i,j} = u_i \cdot v_j$$

$$\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$$
$$p_i = \sum_j u_j \cdot v_{i-j}$$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0

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

Models (Mitchell and Lapata, 2010)

Multiplicative Models

$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v}$

Instances

$\mathbf{p} = \mathbf{u} \odot \mathbf{v}$
$p_i = u_i v_i$

$\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$
$p_{i,j} = u_i \cdot v_j$

$\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$
$p_i = \sum_j u_j \cdot v_{i-j}$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0

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

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

Models (Mitchell and Lapata, 2010)

Multiplicative Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v}$$

Instances

$$\mathbf{p} = \mathbf{u} \odot \mathbf{v}$$
$$p_i = u_i v_i$$

$$\mathbf{p} = \mathbf{u} \otimes \mathbf{v}$$
$$p_{i,j} = u_i \cdot v_j$$

$$\mathbf{p} = \mathbf{u} \circledast \mathbf{v}$$
$$p_i = \sum_j u_j \cdot v_{i-j}$$

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0

practical $\odot$ **difficulty** $=$ [0 48 8 40 0]

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

practical $\circledast$ **difficulty** $=$ [116 50 66 62 80]

Models (Mitchell and Lapata, 2010)

Dilation Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} = \mathbf{U}\mathbf{v}$$
$$U_{ij} = 0, U_{ii} = u_i$$

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \qquad \mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$$

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

$$\mathbf{p} = (\lambda - \mathbf{1})(\mathbf{u}\cdot\mathbf{v})\mathbf{u} + (\mathbf{u}\cdot\mathbf{u})\mathbf{v}$$

Models (Mitchell and Lapata, 2010)

Dilation Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} = \mathbf{U}\mathbf{v}$$

$$U_{ij} = 0, U_{ii} = u_i$$

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \qquad \mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$$

$$\mathbf{v}^{'} = \lambda\mathbf{x} + \mathbf{y} = (\lambda - \mathbf{1})\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} + \mathbf{v}$$

$$\mathbf{p} = (\lambda - \mathbf{1})(\mathbf{u}\cdot\mathbf{v})\mathbf{u} + (\mathbf{u}\cdot\mathbf{u})\mathbf{v}$$

Models (Mitchell and Lapata, 2010)

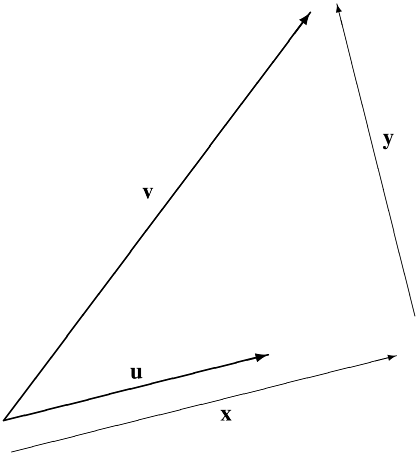
Dilation Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} = \mathbf{U}\mathbf{v}$$
$$U_{ij} = 0, U_{ii} = u_i$$

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \qquad \mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$$

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

$$\mathbf{p} = (\lambda - 1)(\mathbf{u}\cdot\mathbf{v})\mathbf{u} + (\mathbf{u}\cdot\mathbf{u})\mathbf{v}$$



Models (Mitchell and Lapata, 2010)

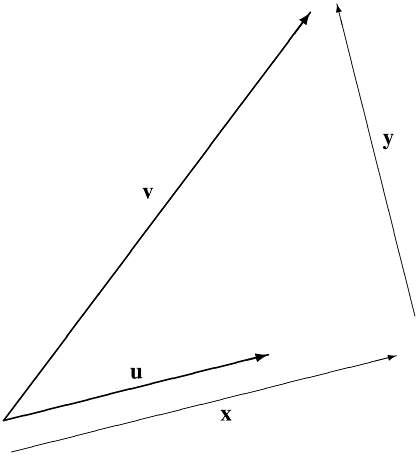
Dilation Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} = \mathbf{U}\mathbf{v}$$
$$U_{ij} = 0, U_{ii} = u_i$$

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \qquad \mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$$

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

$$\mathbf{p} = (\lambda - 1)(\mathbf{u}\cdot\mathbf{v})\mathbf{u} + (\mathbf{u}\cdot\mathbf{u})\mathbf{v}$$



Models (Mitchell and Lapata, 2010)

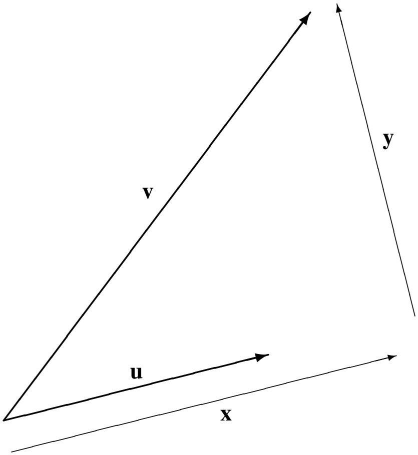
Dilation Models

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} = \mathbf{U}\mathbf{v}$$
$$U_{ij} = 0, U_{ii} = u_i$$

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \qquad \mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$$

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

$$\mathbf{p} = (\lambda - 1)(\mathbf{u}\cdot\mathbf{v})\mathbf{u} + (\mathbf{u}\cdot\mathbf{u})\mathbf{v}$$



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.

	High	**Medium**	**Low**
old person			
kitchen door			
produce effect			

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.

	High	**Medium**	**Low**
old person	elderly lady	right hand	small house
kitchen door			
produce effect			

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.

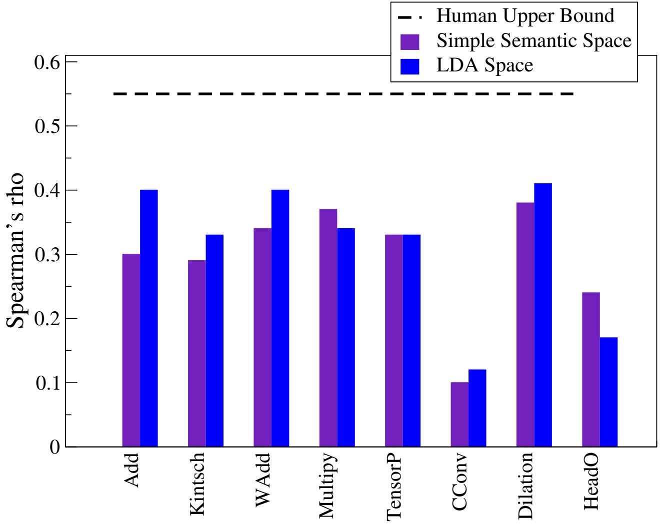
	High	**Medium**	**Low**
old person	elderly lady	right hand	small house
kitchen door	bedroom window	office worker	housing department
produce effect			

Phrase Similarity Task (Kintsch, 2002)

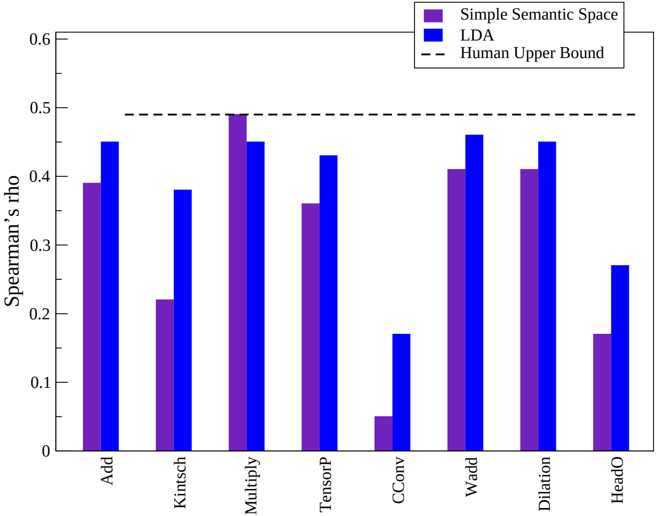
- Elicit similarity judgments for adjective-noun, noun-noun, verb-object combinations.
- Phrase pairs from three bands: High, Medium, Low.
- Compute vectors for phrases, measure their similarity.
- Correlate model similarities with human ratings.

	High	**Medium**	**Low**
old person	elderly lady	right hand	small house
kitchen door	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)

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 $S1 = (w_1 \ldots w_m)$ and $S2 = (w_1 \ldots w_n)$

Task: Classify whether $S1$ and $S2$ are paraphrases or not

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

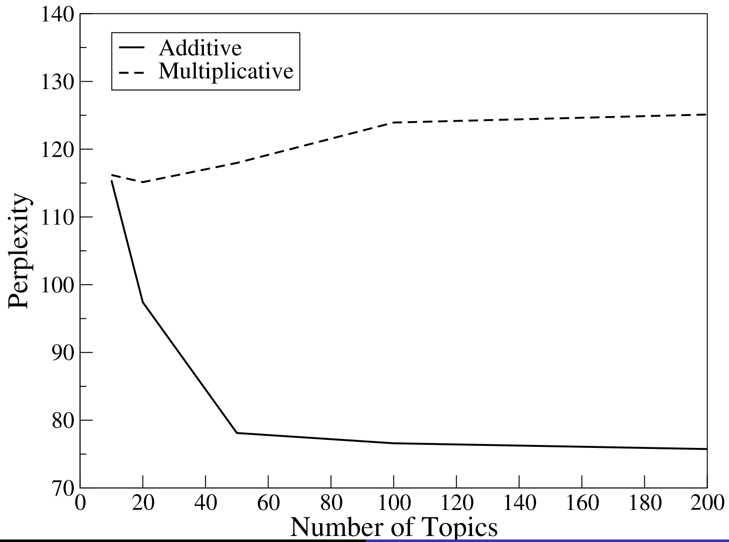
> Referring to him as only “the witness”, Amrozi accused his brother of deliberately distorting his evidence.

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

Model	Acc.	F1
Baseline	66.5	79.9
Mihalcea et al. (2006)	70.3	81.3
Rus et al. (2008)	70.6	80.5
Qiu et al. (2006)	72.0	81.6
Islam et al. (2007)	72.6	81.3
Mitchell and Lapata (2010)	**73.0**	**83.3**
Fernando et al. (2008)	74.1	82.4
Wan et al. (2006)	75.6	83.0
Socher et al. (76.4)	76.4	83.6

LDA Topics


Additive
Multiplicative
Perplexity
Number of Topics
140
130
120
110
100
90
80
70
0
20
40
60
80
100
120
140
160
180
200