Distributional Semantics

Pawan Goyal

CSE, IIT Kharagpur

ACM Summer School

Introduction

What is Semantics?

Introduction

What is Semantics?

The study of meaning: Relation between symbols and their denotata.

Introduction

What is Semantics?

The study of meaning: Relation between symbols and their denotata.
John told Mary that the train moved out of the station at 3 o'clock.

Computational Semantics

Computational Semantics

Computational Semantics

The study of how to automate the process of constructing and reasoning with meaning representations of natural language expressions.

Computational Semantics

Computational Semantics

The study of how to automate the process of constructing and reasoning with meaning representations of natural language expressions.

Methods in Computational Semantics generally fall in two categories:

- **Formal Semantics:** Construction of precise mathematical models of the relations between expressions in a natural language and the world.

Computational Semantics

The study of how to automate the process of constructing and reasoning with meaning representations of natural language expressions.

Methods in Computational Semantics generally fall in two categories:

- **Formal Semantics:** Construction of precise mathematical models of the relations between expressions in a natural language and the world.
  John chases a bat $\rightarrow \exists x[bat(x) \wedge chase(john, x)]$

Computational Semantics

Computational Semantics

The study of how to automate the process of constructing and reasoning with meaning representations of natural language expressions.

Methods in Computational Semantics generally fall in two categories:

- **Formal Semantics:** Construction of precise mathematical models of the relations between expressions in a natural language and the world.
  John chases a bat $\rightarrow \exists x[bat(x) \wedge chase(john, x)]$
- **Distributional Semantics:** The study of statistical patterns of human word usage to extract semantics.

Distributional Hypothesis

Distributional Hypothesis

Distributional Hypothesis: Basic Intuition

"The meaning of a word is its use in language." (Wittgenstein, 1953)

"You know a word by the company it keeps." (Firth, 1957)

Distributional Hypothesis

Distributional Hypothesis: Basic Intuition

"The meaning of a word is its use in language." (Wittgenstein, 1953)

"You know a word by the company it keeps." (Firth, 1957)

$\rightarrow$ Word meaning (whatever it might be) is reflected in linguistic distributions.

Distributional Hypothesis

Distributional Hypothesis: Basic Intuition

"The meaning of a word is its use in language." (Wittgenstein, 1953)

"You know a word by the company it keeps." (Firth, 1957)

$\rightarrow$ Word meaning (whatever it might be) is reflected in linguistic distributions.

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

Distributional Hypothesis

Distributional Hypothesis: Basic Intuition

"The meaning of a word is its use in language." (Wittgenstein, 1953)

"You know a word by the company it keeps." (Firth, 1957)

→ Word meaning (whatever it might be) is reflected in linguistic distributions.

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

→ Semantically similar words tend to have similar distributional patterns.

Distributional Semantics: a linguistic perspective

"If linguistics is to deal with meaning, it can only do so through distributional analysis." (Zellig Harris)

Distributional Semantics: a linguistic perspective

"If linguistics is to deal with meaning, it can only do so through distributional analysis." (Zellig Harris)

"If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution." (Zellig Harris, "Distributional Structure")

Distributional Semantics: a linguistic perspective

"If linguistics is to deal with meaning, it can only do so through distributional analysis." (Zellig Harris)

"If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution." (Zellig Harris, "Distributional Structure")

Differential and not *referential*

Distributional Semantics: a cognitive perspective

Contextual representation

A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various linguistic contexts.

Distributional Semantics: a cognitive perspective

Contextual representation

A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various linguistic contexts.

We learn new words based on contextual cues

Distributional Semantics: a cognitive perspective

Contextual representation

A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various linguistic contexts.

We learn new words based on contextual cues

He filled the **wampimuk** with the substance, passed it around and we all drunk some.

Contextual representation

A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various linguistic contexts.

We learn new words based on contextual cues

He filled the **wampimuk** with the substance, passed it around and we all drunk some.
We found a little **wampimuk** sleeping behind the tree.

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic repesentations from corpus data

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic repesentations from corpus data
- DSMs are models for semantic representations
  - The semantic content is represented by a vector
  - Vectors are obtained through the statistical analysis of the linguistic contexts of a word

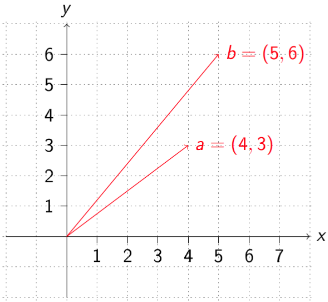
Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic repesentations from corpus data
- DSMs are models for semantic representations
  - The semantic content is represented by a vector
  - Vectors are obtained through the statistical analysis of the linguistic contexts of a word
- Alternative names
  - corpus-based semantics
  - statistical semantics
  - geometrical models of meaning
  - vector semantics
  - word space models

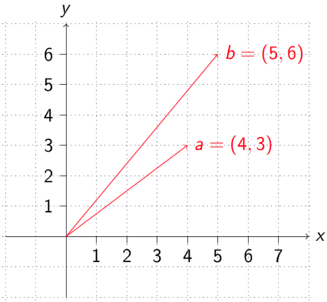
Distributional Semantics: The general intuition

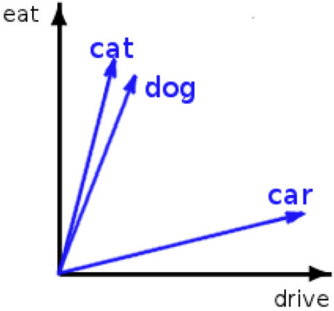
- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude and a direction.
- The **semantic space** has dimensions which correspond to possible contexts, as gathered from a given corpus.

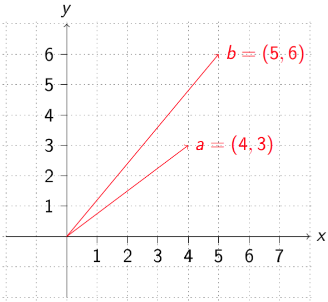
Vector Space

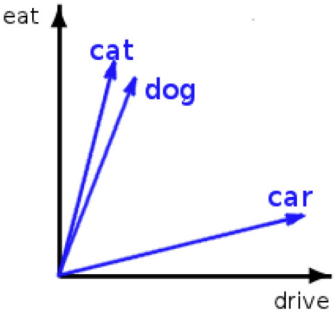


Vector Space









In practice, many more dimensions are used.

$cat =$ [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2,...]

Word Space

Small Dataset

An automobile is a wheeled motor vehicle used for transporting passengers .
A car is a form of transport , usually with four wheels and the capacity to carry around five passengers .
Transport for the London games is limited , with spectators strongly advised to avoid the use of cars .
The London 2012 soccer tournament began yesterday , with plenty of goals in the opening matches .
Giggs scored the first goal of the football tournament at Wembley , North London .
Bellamy was largely a passenger in the football match , playing no part in either goal .

Target words: ⟨automobile, car, soccer, football⟩
Term vocabulary: ⟨wheel, transport, passenger, tournament, London, goal, match⟩

Constructing Word spaces

Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**

Constructing Word spaces

Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**
- Define a **context window**, number of words surrounding target word

Constructing Word spaces

Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**
- Define a **context window**, number of words surrounding target word
    - The context can in general be defined in terms of documents, paragraphs or sentences.

Constructing Word spaces

Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**
- Define a **context window**, number of words surrounding target word
  - The context can in general be defined in terms of documents, paragraphs or sentences.
- Count number of times the target word co-occurs with the context words: **co-occurrence matrix**

Constructing Word spaces

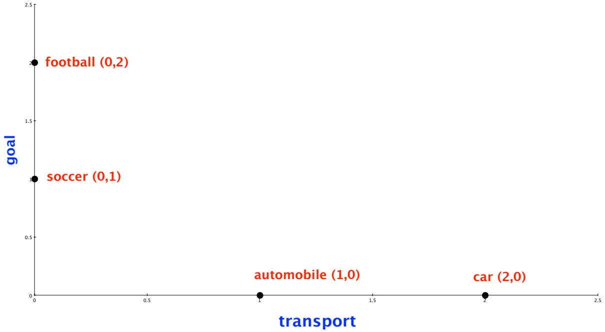
Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**
- Define a **context window**, number of words surrounding target word
  - The context can in general be defined in terms of documents, paragraphs or sentences.
- Count number of times the target word co-occurs with the context words: **co-occurrence matrix**
- Build vectors out of (a function of) these co-occurrence counts

Constructing Word spaces: distributional vectors

distributional matrix = targets X contexts

	wheel	transport	passenger	tournament	London	goal	match
tomobile	1	1	1	0	0	0	0
r	1	2	1	0	1	0	0
ccer	0	0	0	1	1	1	1
otball	0	0	1	1	1	2	1


football (0,2)
soccer (0,1)
automobile (1,0)
car (2,0)
goal
transport
2.5
1.5
0.5
0
0.5
1.5
2.5

Computing similarity

	wheel	transport	passenger	tournament	London	goal	match
automobile	1	1	1	0	0	0	0
car	1	2	1	0	1	0	0
soccer	0	0	0	1	1	1	1
football	0	0	1	1	1	2	1

Using simple vector product

automobile . car = 4

automobile . soccer = 0

automobile . football = 1

car . soccer = 1

car . football = 2

soccer . football = 5

Vector Space Model without distributional similarity

Words are treated as atomic symbols

Vector Space Model without distributional similarity

Words are treated as atomic symbols

One-hot representation

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Distributional Similarity Based Representations

You know a word by the company it keeps

Distributional Similarity Based Representations

You know a word by the company it keeps

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

Distributional Similarity Based Representations

You know a word by the company it keeps

government debt problems turning into banking crises as has happened in

saying that Europe needs unified banking regulation to replace the hodgepodge

These words will represent banking

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

$\Downarrow$

Weight the contexts (optional)

$\Downarrow$

Build the distributional matrix

$\Downarrow$

Reduce the matrix dimensions (optional)

$\Downarrow$

Compute the vector distances on the (reduced) matrix

Many design choices

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

Many design choices

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

General Questions

- How do the rows (words, ...) relate to each other?
- How do the columns (contexts, documents, ...) relate to each other?

The parameter space

A number of parameters to be fixed

- Which type of context?
- Which weighting scheme?
- Which similarity measure?
- ...

A specific parameter setting determines a particular type of DSM (e.g. LSA, HAL, etc.)

Documents as context: Word $\times$ document

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Words as context: Word × Word

	against	age	agent	ages	ago	agree	ahead	ain.t	air	aka	al
against	2003	90	39	20	88	57	33	15	58	22	24
age	90	1492	14	39	71	38	12	4	18	4	39
agent	39	14	507	2	21	5	10	3	9	8	25
ages	20	39	2	290	32	5	4	3	6	1	6
ago	88	71	21	32	1164	37	25	11	34	11	38
agree	57	38	5	5	37	627	12	2	16	19	14
ahead	33	12	10	4	25	12	429	4	12	10	7
ain't	15	4	3	3	11	2	4	166	0	3	3
air	58	18	9	6	34	16	12	0	746	5	11
aka	22	4	8	1	11	19	10	3	5	261	9
al	24	39	25	6	38	14	7	3	11	9	861

Words as contexts

Parameters

- Window size
- Window shape - rectangular/triangular/other

Words as contexts

Parameters

- Window size
- Window shape - rectangular/triangular/other

Consider the following passage

Suspected communist rebels on 4 July 1989 killed Col. Herminio Taylo, police chief of Makati, the Philippines major financial center, in an escalation of street violence sweeping the Capitol area. The gunmen shouted references to the rebel New People's Army. They fled in a commandeered passenger jeep. The military says communist rebels have killed up to 65 soldiers and police in the Capitol region since January.

Words as contexts

Parameters

- Window size
- Window shape - rectangular/triangular/other

5 words window (unfiltered): 2 words either side of the target word

Suspected communist rebels on 4 July 1989 killed Col. Herminio Taylo, police chief of Makati, the Philippines major financial center, in an escalation of street violence sweeping the Capitol area. The gunmen shouted references to the rebel New People's Army. They fled in a commandeered passenger jeep. The military says communist rebels have killed up to 65 soldiers and police in the Capitol region since January.

Words as contexts

Parameters

- Window size
- Window shape - rectangular/triangular/other

5 words window (filtered): 2 words either side of the target word

Suspected communist rebels on 4 July 1989 killed Col. Herminio Taylo, police chief of Makati, the Philippines major financial center, in an escalation of street violence sweeping the Capitol area. The gunmen shouted references to the rebel New People's Army. They fled in a commandeered passenger jeep. The military says communist rebels have killed up to 65 soldiers and police in the Capitol region since January.

Context weighting: documents as context

Context weighting: documents as context

Indexing function F: Essential factors

- **Word frequency (f_{ij}):** How many times a word appears in the document? $F \propto f_{ij}$
- **Document length ($|D_i|$):** How many words appear in the document? $F \propto \frac{1}{|D_i|}$
- **Document frequency (N_j):** Number of documents in which a word appears. $F \propto \frac{1}{N_j}$

Context weighting: documents as context

Indexing function F: Essential factors

- **Word frequency (f_{ij}):** How many times a word appears in the document? $F \propto f_{ij}$
- **Document length ($|D_i|$):** How many words appear in the document? $F \propto \frac{1}{|D_i|}$
- **Document frequency (N_j):** Number of documents in which a word appears. $F \propto \frac{1}{N_j}$

Indexing Weight: tf-Idf

- $f_{ij} * log(\frac{N}{N_j})$ for each term, normalize the weight in a document with respect to L_2-norm.

Context weighting: words as context

Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Association measures are used to give more weight to contexts that are more significantly associted with a targer word.

Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Association measures are used to give more weight to contexts that are more significantly associted with a targer word.

- The less frequent the target and context element are, the higher the weight given to their co-occurrence count should be.

Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Association measures are used to give more weight to contexts that are more significantly associted with a targer word.

- The less frequent the target and context element are, the higher the weight given to their co-occurrence count should be.
  $\Rightarrow$ Co-occurrence with frequent context element *small* is less informative than co-occurrence with rarer *domesticated*.

Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Association measures are used to give more weight to contexts that are more significantly associted with a targer word.

- The less frequent the target and context element are, the higher the weight given to their co-occurrence count should be.
  $\Rightarrow$ Co-occurrence with frequent context element *small* is less informative than co-occurrence with rarer *domesticated*.
- different measures - e.g., Mutual information, Log-likelihood ratio

Pointwise Mutual Information (PMI)

$$PMI(w_1, w_2) = log_2 \frac{P_{corpus}(w_1, w_2)}{P_{ind}(w_1, w_2)}$$

Pointwise Mutual Information (PMI)

$$PMI(w_1,w_2) = log_2 \frac{P_{corpus}(w_1,w_2)}{P_{ind}(w_1,w_2)}$$

$$PMI(w_1,w_2) = log_2 \frac{P_{corpus}(w_1,w_2)}{P_{corpus}(w_1)P_{corpus}(w_2)}$$

Pointwise Mutual Information (PMI)

$$PMI(w_1, w_2) = log_2 \frac{P_{corpus}(w_1, w_2)}{P_{ind}(w_1, w_2)}$$

$$PMI(w_1, w_2) = log_2 \frac{P_{corpus}(w_1, w_2)}{P_{corpus}(w_1)P_{corpus}(w_2)}$$

$$P_{corpus}(w_1, w_2) = \frac{freq(w_1, w_2)}{N}$$

$$P_{corpus}(w) = \frac{freq(w)}{N}$$

PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

Bias towards infrequent events

Consider w_j having the maximum association with w_i,
$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$

PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

Bias towards infrequent events

Consider w_j having the maximum association with w_i,
$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$
PMI increases as the probability of w_i decreases.

PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

Bias towards infrequent events

Consider w_j having the maximum association with w_i,
$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$
PMI increases as the probability of w_i decreases.
Also, consider a word w_j that occurs once in the corpus, also in the context of w_i.

PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

Bias towards infrequent events

Consider w_j having the maximum association with w_i,
$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$
PMI increases as the probability of w_i decreases.
Also, consider a word w_j that occurs once in the corpus, also in the context of w_i. A discounting factor proposed by Pantel and Lin:

$$\delta_{ij} = \frac{f_{ij}}{f_{ij}+1} \frac{min(f_i, f_j)}{min(f_i, f_j)+1}$$

$PMI_{new}(w_i, w_j) = \delta_{ij} PMI(w_i, w_j)$

Distributional Vectors: Example

Normalized Distributional Vectors using Pointwise Mutual Information

petroleum	oil:0.032 gas:0.029 crude:0.029 barrels:0.028 exploration:0.027 barrel:0.026 opec:0.026 refining:0.026 gasoline:0.026 fuel:0.025 natural:0.025 exporting:0.025
drug	trafficking:0.029 cocaine:0.028 narcotics:0.027 fda:0.026 police:0.026 abuse:0.026 marijuana:0.025 crime:0.025 colombian:0.025 arrested:0.025 addicts:0.024
insurance	insurers:0.028 premiums:0.028 lloyds:0.026 reinsurance:0.026 underwriting:0.025 pension:0.025 mortgage:0.025 credit:0.025 investors:0.024 claims:0.024 benefits:0.024
forest	timber:0.028 trees:0.027 land:0.027 forestry:0.026 environmental:0.026 species:0.026 wildlife:0.026 habitat:0.025 tree:0.025 mountain:0.025 river:0.025 lake:0.025
robotics	robots:0.032 automation:0.029 technology:0.028 engineering:0.026 systems:0.026 sensors:0.025 welding:0.025 computer:0.025 manufacturing:0.025 automated:0.025

Application to Query Expansion: Addressing Term Mismatch

Term Mismatch Problem in Information Retrieval

- Stems from the word independence assumption during document indexing.
- User query: *insurance cover which pays for long term care.*
- A relevant document may contain terms different from the actual user query.
- Some relevant words concerning this query: $\{medicare, premiums, insurers\}$

Application to Query Expansion: Addressing Term Mismatch

Term Mismatch Problem in Information Retrieval

- Stems from the word independence assumption during document indexing.
- User query: *insurance cover which pays for long term care.*
- A relevant document may contain terms different from the actual user query.
- Some relevant words concerning this query: $\{medicare, premiums, insurers\}$

Using DSMs for Query Expansion

Given a user query, reformulate it using related terms to enhance the retrieval performance.

- The distributional vectors for the query terms are computed.
- Expanded query is obtained by a linear combination or a functional combination of these vectors.

Query Expansion using Unstructured DSMs

TREC Topic 104: *catastrophic health insurance*

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

Query Expansion using Unstructured DSMs

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** ...
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

Query Expansion using Unstructured DSMs

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** ...
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

TREC Topic 355: ocean remote sensing

Query Representation: radiometer:1.0 landsat:0.97 ionosphere:0.94 cnes:0.84 altimeter:0.83 nasda:0.81 meterology:0.81 cartography:0.78 geostationary:0.78 doppler:0.78 oceanographic:0.76

Query Expansion using Unstructured DSMs

TREC Topic 104: catastrophic health insurance

Query Representation: surtax:1.0 hcfa:0.97 medicare:0.93 hmos:0.83 medicaid:0.8 hmo:0.78 beneficiaries:0.75 ambulatory:0.72 premiums:0.72 hospitalization:0.71 hhs:0.7 reimbursable:0.7 deductible:0.69

- Broad expansion terms: **medicare, beneficiaries, premiums** ...
- Specific domain terms: **HCFA** (Health Care Financing Administration), **HMO** (Health Maintenance Organization), **HHS** (Health and Human Services)

TREC Topic 355: ocean remote sensing

Query Representation: radiometer:1.0 landsat:0.97 ionosphere:0.94 cnes:0.84 altimeter:0.83 nasda:0.81 meterology:0.81 cartography:0.78 geostationary:0.78 doppler:0.78 oceanographic:0.76

- Broad expansion terms: **radiometer, landsat, ionosphere** ...
- Specific domain terms: **CNES** (Centre National dÉtudes Spatiales) and **NASDA** (National Space Development Agency of Japan)

Similarity Measures for Binary Vectors

Let X and Y denote the binary distributional vectors for words X and Y.

Similarity Measures

Dice coefficient : $\frac{2|X \cap Y|}{|X|+|Y|}$

Similarity Measures for Binary Vectors

Let X and Y denote the binary distributional vectors for words X and Y.

Similarity Measures

$$\text{Dice coefficient : } \frac{2|X \cap Y|}{|X| + |Y|}$$

$$\text{Jaccard Coefficient : } \frac{|X \cap Y|}{|X \cup Y|}$$

Similarity Measures for Binary Vectors

Let X and Y denote the binary distributional vectors for words X and Y.

Similarity Measures

$$\text{Dice coefficient : } \frac{2|X \cap Y|}{|X| + |Y|}$$

$$\text{Jaccard Coefficient : } \frac{|X \cap Y|}{|X \cup Y|}$$

$$\text{Overlap Coefficient : } \frac{|X \cap Y|}{min(|X|, |Y|)}$$

Similarity Measures for Binary Vectors

Let X and Y denote the binary distributional vectors for words X and Y.

Similarity Measures

$$\text{Dice coefficient : } \frac{2|X \cap Y|}{|X|+|Y|}$$

$$\text{Jaccard Coefficient : } \frac{|X \cap Y|}{|X \cup Y|}$$

$$\text{Overlap Coefficient : } \frac{|X \cap Y|}{min(|X|,|Y|)}$$

Jaccard coefficient penalizes small number of shared entries, while Overlap coefficient uses the concept of inclusion.

Similarity Measures for Vector Spaces

Let $\vec{X}$ and $\vec{Y}$ denote the distributional vectors for words X and Y.
$\vec{X} = [x_1, x_2, \ldots, x_n]$, $\vec{Y} = [y_1, y_2, \ldots, y_n]$

Let $\vec{X}$ and $\vec{Y}$ denote the distributional vectors for words X and Y.
$\vec{X} = [x_1, x_2, \ldots, x_n]$, $\vec{Y} = [y_1, y_2, \ldots, y_n]$

Similarity Measures

Cosine similarity : $cos(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}||\vec{Y}|}$

Similarity Measures for Vector Spaces

Let $\vec{X}$ and $\vec{Y}$ denote the distributional vectors for words X and Y.
$\vec{X} = [x_1, x_2, \ldots, x_n]$, $\vec{Y} = [y_1, y_2, \ldots, y_n]$

Similarity Measures

$$\text{Cosine similarity : } cos(\vec{X}, \vec{Y}) = \frac{\bar{X} \cdot \vec{Y}}{|\vec{X}||\vec{Y}|}$$

$$\text{Euclidean distance : } |\vec{X} - \vec{Y}| = \sqrt{\Sigma_{i=1}^{n} (x_i - y_i)^2}$$

Similarity Measure for Probability Distributions

Let p and q denote the probability distributions corresponding to two distributional vectors.

Similarity Measure for Probability Distributions

Let p and q denote the probability distributions corresponding to two distributional vectors.

Similarity Measures

KL-divergence : $D(p||q) = \Sigma_i p_i log \frac{p_i}{q_i}$

Information Radius : $D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$

L_1-norm : $\Sigma_i |p_i - q_i|$

Dimensionality Reduction

- Reduce the target-word by context matrix to a lower dimensionality matrix
- Two main reasons:
  - efficiency - sometimes the marix is so large that you don't want to construct it explicitly.

Dimensionality Reduction

- Reduce the target-word by context matrix to a lower dimensionality matrix
- Two main reasons:
  - efficiency - sometimes the marix is so large that you don't want to construct it explicitly.
  - smoothing - capture "latent dimensions" that generalize over sparser surface dimensions, synonym vectors may not be orthogonal.

Latent Semantic Indexing

- General technique from Linear Algebra (similar to Principal Component Analysis, PCA)
- Given a matrix (e.g., a word-by-document matrix) of dimensionality $m \times n$ of rank l, construct a rank k model ($k << l$) with the best possible least squares fit
- The reduced matrix should preserve most of the variance in the original matrix.

Latent Semantic Indexing

The Singular Value Decomposition (SVD) of an m-by-n matrix A is:

$$A = U\Sigma V^T$$

- U is an $m \times l$ matrix, V is an $n \times l$ matrix, and Σ is an $l \times l$ matrix, where l is the rank of the matrix A.
- The $m-$dimensional vectors making up the columns of U are called left singular vectors.
- The n-dimensional vectors making up the columns of V are called right singular vectors.
- The values on the diagonal of Σ are called the singular values.
- Latent Semantic Indexing

$$A_k = U_k \Sigma_k {V_k}^T$$

SVD: An Example

Sample dataset: titles of nine technical memoranda

c1: Human machine interface for ABC computer applications
c2: A survey of user opinion of computer system response time
c3: The EPS user interface management system
c4: System and human system engineering testing of EPS
c5: Relation of user perceived response time to error measurement
m1: The generation of random, binary, ordered trees
m2: The intersection graph of paths in trees
m3: Graph minors IV: Widths of trees and well-quasi-ordering
m4: Graph minors: A survey

SVD: An Example

Sim(human,user) = 0.0, *Sim(human,minors)* = 0.0

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

SVD: An Example

$U =$

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

SVD: An Example

$$\Sigma = \begin{bmatrix} 3.34 & & & & & & & & \\ & 2.54 & & & & & & & \\ & & 2.35 & & & & & & \\ & & & 1.64 & & & & & \\ & & & & 1.50 & & & & \\ & & & & & 1.31 & & & \\ & & & & & & 0.85 & & \\ & & & & & & & 0.56 & \\ & & & & & & & & 0.36 \end{bmatrix}$$

SVD: An Example

$V =$

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

SVD: An Example

Sim(human, user) = 0.94, *Sim(human, minors)* = -0.83

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Attributional Similarity vs. Relational Similarity

Attributional Similarity

The attributional similarity between two words a and b depends on the degree of correspondence between the properties of a and b.

Ex: dog and wolf

Relational Similarity

Two pairs(a, b) and (c, d) are relationally similar if they have many similar relations.

Ex: dog: bark and cat: meow

Relational Similarity: Pair-pattern matrix

Pair-pattern matrix

- Row vectors correspond to pairs of words, such as *mason: stone* and *carpenter: wood*
- Column vectors correspond to the patterns in which the pairs occur, e.g. *X cuts Y* and *X works with Y*
- Compute the similarity of rows to find similar pairs

Relational Similarity: Pair-pattern matrix

Pair-pattern matrix

- Row vectors correspond to pairs of words, such as *mason: stone* and *carpenter: wood*
- Column vectors correspond to the patterns in which the pairs occur, e.g. *X cuts Y* and *X works with Y*
- Compute the similarity of rows to find similar pairs

Extended Distributional Hypothesis; Lin and Pantel

Patterns that co-occur with similar pairs tend to have similar meanings.

Relational Similarity: Pair-pattern matrix

Pair-pattern matrix

- Row vectors correspond to pairs of words, such as *mason: stone* and *carpenter: wood*
- Column vectors correspond to the patterns in which the pairs occur, e.g. *X cuts Y* and *X works with Y*
- Compute the similarity of rows to find similar pairs

Extended Distributional Hypothesis; Lin and Pantel

Patterns that co-occur with similar pairs tend to have similar meanings.
This matrix can also be used to measure the semantic similarity of patterns.

Relational Similarity: Pair-pattern matrix

Pair-pattern matrix

- Row vectors correspond to pairs of words, such as *mason: stone* and *carpenter: wood*
- Column vectors correspond to the patterns in which the pairs occur, e.g. *X cuts Y* and *X works with Y*
- Compute the similarity of rows to find similar pairs

Extended Distributional Hypothesis; Lin and Pantel

Patterns that co-occur with similar pairs tend to have similar meanings.
This matrix can also be used to measure the semantic similarity of patterns.
Given a pattern such as “X solves Y”, you can use this matrix to find similar patterns, such as “Y is solved by X”, “Y is resolved in X”, “X resolves Y”.

Structured DSMs

Basic Issue

- Words may not be the basic context units anymore
- How to capture and represent syntactic information?
  X solves Y and *Y is solved by X*

Structured DSMs

Basic Issue

- Words may not be the basic context units anymore
- How to capture and represent syntactic information?
  X solves Y and *Y is solved by X*

An Ideal Formalism

- Should mirror semantic relationships as close as possible
- Incorporate word-based information and syntactic analysis
- Should be applicable to different languages

Structured DSMs

Basic Issue

- Words may not be the basic context units anymore
- How to capture and represent syntactic information?
  X solves Y and *Y is solved by X*

An Ideal Formalism

- Should mirror semantic relationships as close as possible
- Incorporate word-based information and syntactic analysis
- Should be applicable to different languages

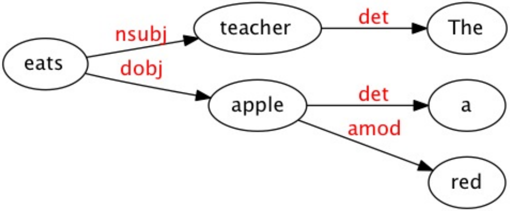
Use Dependency grammar framework

Structured DSMs

Structured DSMs

Using Dependency Structure: How does it help?

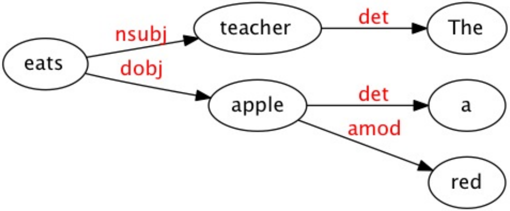
The teacher eats a red apple.



Structured DSMs

Using Dependency Structure: How does it help?

The teacher eats a red apple.



- 'teacher' is not a legitimate context for 'red'.
- The 'object' relation connecting 'eat' and 'apple' is treated as a different type of co-occurrence from the 'modifier' relation linking 'red' and 'apple'.

Structured DSMs

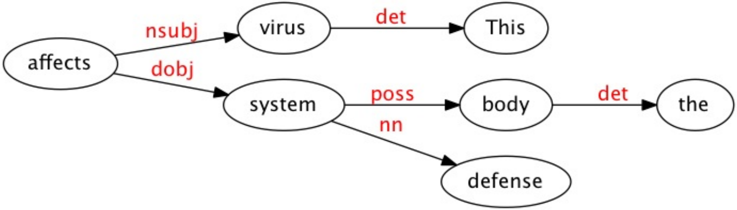
Structured DSMs: Words as 'legitimate' contexts

- Co-occurrence statistics are collected using parser-extracted relations.
- To qualify as context of a target item, a word must be linked to it by some (interesting) lexico-syntactic relation

Structured DSMs

Distributional models, as guided by dependency

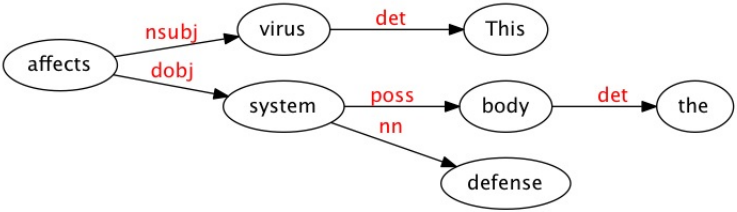
Ex: For the sentence '*This virus affects the body's defense system.*', the dependency parse is:



Structured DSMs

Distributional models, as guided by dependency

Ex: For the sentence '*This virus affects the body's defense system.*', the dependency parse is:



Word vectors

<system, dobj, affects> ...

Corpus-derived ternary data can also be mapped onto a 2-way matrix

2-way matrix

<system, dobj, affects>
<virus, nsubj, affects>

The dependency information can be dropped

- <system, dobj, affects> ⇒ <system, affects>
- <virus, nsubj, affects> ⇒ <virus, affects>

2-way matrix

<system, dobj, affects>
<virus, nsubj, affects>

The dependency information can be dropped

- <system, dobj, affects> $\Rightarrow$ <system, affects>
- <virus, nsubj, affects> $\Rightarrow$ <virus, affects>

Link and one word can be concatenated and treated as attributes

- *virus*={nsubj-affects:0.05,...},
- *system*={dobj-affects:0.03,...}

Structured DSMs for Selectional Preferences

Selectional Preferences for Verbs

Most verbs prefer arguments of a particular type. This regularity is known as selectional preference.

Structured DSMs for Selectional Preferences

Selectional Preferences for Verbs

Most verbs prefer arguments of a particular type. This regularity is known as selectional preference.

- From a parsed corpus, noun vectors are calculated as shown for ‘virus’ and ‘system’.

Structured DSMs for Selectional Preferences

Selectional Preferences for Verbs

Most verbs prefer arguments of a particular type. This regularity is known as selectional preference.

- From a parsed corpus, noun vectors are calculated as shown for 'virus' and 'system'.

	obj-carry	obj-buy	obj-drive	obj-eat	obj-store	sub-fly	...
car	0.1	0.4	0.8	0.02	0.2	0.05	...
vegetable	0.3	0.5	0	0.6	0.3	0.05	...
biscuit	0.4	0.4	0	0.5	0.4	0.02	...
...	...	...	...	...	...	...	...

Structured DSMs for Selectional Preferences

Selectional Preferences

- Suppose we want to compute the selectional preferences of the nouns as object of verb 'eat'.

Structured DSMs for Selectional Preferences

Selectional Preferences

- Suppose we want to compute the selectional preferences of the nouns as object of verb 'eat'.
- n nouns having highest weight in the dimension 'obj-eat' are selected, let $\{$vegetable, biscuit,$\ldots\}$ be the set of these n nouns.

Structured DSMs for Selectional Preferences

Selectional Preferences

- Suppose we want to compute the selectional preferences of the nouns as object of verb ‘eat’.
- n nouns having highest weight in the dimension ‘obj-eat’ are selected, let $\{$vegetable, biscuit,$\ldots\}$ be the set of these n nouns.
- The complete vectors of these n nouns are used to obtain an ‘object prototype’ of the verb.

Structured DSMs for Selectional Preferences

Selectional Preferences

- Suppose we want to compute the selectional preferences of the nouns as object of verb 'eat'.
- n nouns having highest weight in the dimension 'obj-eat' are selected, let $\{$vegetable, biscuit,$\ldots\}$ be the set of these n nouns.
- The complete vectors of these n nouns are used to obtain an 'object prototype' of the verb.
- 'object prototype' will indicate various attributes such as these nouns can be consumed, bought, carried, stored etc.

Structured DSMs for Selectional Preferences

Selectional Preferences

- Suppose we want to compute the selectional preferences of the nouns as object of verb 'eat'.
- n nouns having highest weight in the dimension 'obj-eat' are selected, let $\{$vegetable, biscuit,$\ldots\}$ be the set of these n nouns.
- The complete vectors of these n nouns are used to obtain an 'object prototype' of the verb.
- 'object prototype' will indicate various attributes such as these nouns can be consumed, bought, carried, stored etc.
- Similarity of a noun to this 'object prototype' is used to denote the plausibility of that noun being an object of verb 'eat'.

Distributional Memory (DM); Baroni and Lenci (2010)

Distributional Memory (DM): A unified framework

- The core geometrical structure of DM is a 3-way object, a third order tensor.
    - DM represents distributional facts as word-link-word tuples
    - Tuples are formalized as a ternary structure, which can be utilized for a unified model for distributional semantics

Distributional Memory (DM); Baroni and Lenci (2010)

Distributional Memory (DM): A unified framework

- The core geometrical structure of DM is a 3-way object, a third order tensor.
  - DM represents distributional facts as word-link-word tuples
  - Tuples are formalized as a ternary structure, which can be utilized for a unified model for distributional semantics
- Third order tensor can be projected onto 2-way matrices, generating different semantic spaces “on demand”
  - Alternate views of the same underlying distributional object

Weighted tuple structure

- W_1, W_2 : sets of strings representing content words
- L : a set of strings representing syntagmatic co-occurrence links between words
- T : a set of corpus derived tuples $t = < w_1, l, w_2 >$ such that w_1 co-occurs with w_2 and l represents the type of this co-occurrence relation
- v_t : a tuple weight, assigned by a scoring function $\sigma : W_1 \times L \times W_2 \rightarrow R$

Weighted tuple structure

- W_1, W_2 : sets of strings representing content words
- L : a set of strings representing syntagmatic co-occurrence links between words
- T : a set of corpus derived tuples $t =< w_1, l, w_2 >$ such that w_1 co-occurs with w_2 and l represents the type of this co-occurrence relation
- v_t : a tuple weight, assigned by a scoring function $\sigma : W_1 \times L \times W_2 \to R$

Weighted tuple structure

A set T_W of weighted distributional tuples $T_w =< t, v_t >$ for all $t \in T$ and $\sigma(t) = v_t$

Weighted tuple structure

w_1	l	w_2	σ	w_1	l	w_2	σ
marine	own	bomb	40.0	sergeant	use	gun	51.9
marine	use	bomb	82.1	sergeant	own	book	8.0
marine	own	gun	85.3	sergeant	use	book	10.1
marine	use	gun	44.8	teacher	own	bomb	5.2
marine	own	book	3.2	teacher	use	bomb	7.0
marine	use	book	3.3	teacher	own	gun	9.3
sergeant	own	bomb	16.7	teacher	use	gun	4.7
sergeant	use	bomb	69.5	teacher	own	book	48.4
sergeant	own	gun	73.4	teacher	use	book	53.6

Constraints on T_W

- $W_1 = W_2$
- inverse link constraint:
  $<<$marine, use, bomb$>, v_t >\Rightarrow<<$bomb,use^{-1},marine$>, v_t >$

The DM semantic spaces

4 distinc semantic vector spaces

- word by link-word ($W_1 \times LW_2$)
- word-word by link ($W_1W_2 \times L$)
- word-link by word ($W_1L \times W_2$)
- link by word-word ($L \times W_1W_2$)

Experimental Framework

- A corpus containing 2.83 billion tokens
- $W_1 = W_2 = 30693$ (most frequent 20000 nouns, 5000 verbs and 5000 adjectives)

Experimental Framework

- A corpus containing 2.83 billion tokens
- $W_1 = W_2 = 30693$ (most frequent 20000 nouns, 5000 verbs and 5000 adjectives)

Links using dependency information

sbj_intr subject of a verb with no direct object.
The teacher is singing → <teacher, sbj_intr, sing>

sbj_tr subject of a verb that occurs with a direct object.
The soldier is reading a book → <soldier, sbj_tr,read>

obj direct object: *The soldier is reading a book* → <book,obj,read>

iobj indirect object in a double object construction.
The soldier gave the woman a book → <woman, iobj, give>

Experimental Framework

Links using dependency information

nmod noun modifier: *good teacher* $\rightarrow$ <good, nmod, teacher>

coord noun coordination: *teachers and soldiers* $\rightarrow$ <teacher, coord, soldier>

preposition A diferent link for each preposition
I saw a soldier with the gun $\rightarrow$ <gun, with, soldier>

Complex Links

Structure

The tall soldier has already shot $\rightarrow$ $<$soldier, sbj_intr+n-the-j+vn-aux-already, shoot$>$

- pattern+*suffix*
- suffix is formed by two substrings separated by a '+'
- each substring encodes the features of w_1 and w_2: POS, morphology (number, tense), presence of article, adjective, adverb

Complex Links

Structure

The tall soldier has already shot $\rightarrow$ $<$soldier, sbj_intr+n-the-j+vn-aux-already, shoot$>$

- pattern+*suffix*
- suffix is formed by two substrings separated by a '+'
- each substring encodes the features of w_1 and w_2: POS, morphology (number, tense), presence of article, adjective, adverb
- For the above example: 'subj_intr' is the pattern,
- *n-the-j:* w_1 is a singular noun (n), definite (the) and has an adjective (*j*)
- *vn-aux-already:* w_2 is a past-participle (vn), has an auxiliary (aux) and is modified by $already$

Example of complex links

such_as links two nouns in *NOUN such as NOUN* and *such NOUN as NOUN*: *animals such as cats* → <animal, such_as+ns+ns, cat>

as_adj_as links adjective and noun matching *as ADJ as (a|the) NOUN*: *as sharp as a knife* → <sharp, as_adj_as+j+n-a,knife>

attribute_noun 127 nouns extracted from Wordnet expressing attributes of concepts, such as *size, color* or *height*.
Templates: *(the) attribute_noun of (a|the) NOUN is ADJ* and *(a|the) ADJ attribute_noun of NOUN*:
the color of strawberries is red → <red,color+j+ns,strawberry>

Various Modes

A_{mode-1}	1: ⟨own,bomb⟩	2: ⟨use,bomb⟩	3: ⟨own,gun⟩	4: ⟨use,gun⟩	5: ⟨own,book⟩	6: ⟨use,book⟩
1:marine	40.0	82.1	85.3	44.8	3.2	3.3
2:sergeant	16.7	69.5	73.4	51.9	8.0	10.1
3:teacher	5.2	7.0	9.3	4.7	48.4	53.6

B_{mode-2}	1: ⟨marine, bomb⟩	2 ⟨sergeant, bomb⟩	3: ⟨teacher, bomb⟩	4: ⟨marine, gun⟩	5: ⟨sergeant, gun⟩	6: ⟨teacher, gun⟩	7: ⟨marine, book⟩	8: ⟨sergeant, book⟩	9: ⟨teacher, book⟩
1:own	40.0	16.7	5.2	85.3	73.4	9.3	3.2	8.0	48.4
2:use	82.1	69.5	7.0	44.8	51.9	4.7	3.3	10.1	53.6

C_{mode-3}	1: ⟨marine,own⟩	2: ⟨marine,use⟩	3: ⟨sergeant,own⟩	4: ⟨sergeant,use⟩	5: ⟨teacher,own⟩	6: ⟨teacher,use⟩
1:bomb	40.0	82.1	16.7	69.5	5.2	7.0
2:gun	85.3	44.8	73.4	51.9	9.3	4.7
3:book	3.2	3.3	8.0	10.1	48.4	53.6

Some Pair-Problems Addressed

Solving Analogy Problems

Multiple choice questions with one target (*ostrich-bird*) and five candidate analogies (*lion-cat, goose-flock, ewe-sheep, cub-bear, primate-monkey*)

Some Pair-Problems Addressed

Solving Analogy Problems

Multiple choice questions with one target (*ostrich-bird*) and five candidate analogies (*lion-cat, goose-flock, ewe-sheep, cub-bear, primate-monkey*)

Relation Classification

Cause-Effect	cycling-happiness
Purpose	album-picture
Location-At	pain-chest
Time-At	snack-midnight

Google Syntactic n-grams

...but sequential context is only a proxy

(often misleading)

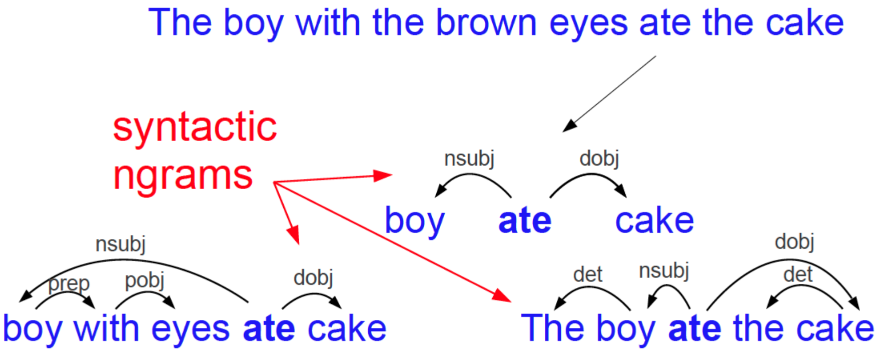
The boy with the brown eyes ate the cake

eyes, **ate**, the

brown, eyes, **ate**, the, cake

Google Syntactic n-grams

what we really care for is the **syntactic context**



Dataset

English Google Books

~3.5M books
published between 1520 to 2008
(most after 1800)

~350B words
~x100 times larger than prev efforts

Encoding Syntactic ngrams

verbargs:

```
covering     hands/NNS/nsubj/2 covering/VBG/dep/0 her/PRP$/poss/4 face/NN/dobj/2  106
covers  as/IN/mark/3 water/NN/nsubj/3 covers/VBZ/advcl/0 the/DT/det/5 sea/NN/dobj/3
126
```

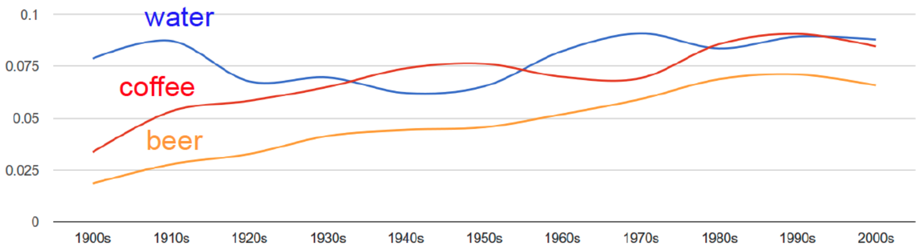
Encoding Syntactic ngrams

verbargs:

```
covering      hands/NNS/nsubj/2 covering/VBG/dep/0 her/PRP$/poss/4 face/NN/dobj/2  106
covers  as/IN/mark/3 water/NN/nsubj/3 covers/VBZ/advcl/0 the/DT/det/5 sea/NN/dobj/3
126
```

```
cease   cease/VB/ccomp/0 for/IN/prep/1 an/DT/det/4 instant/NN/pobj/2
56    1834,2      1835,1      1856,1      1863,1      1871,1      1872,1
1874,1      1875,3      1880,2      1883,2      1889,1      1904,7
1905,2      1915,5      1918,1      1961,1      1963,5      1973,2
1975,1      1977,1      1981,2      1987,2      1988,1      1989,1
1991,1      1996,5      2000,1      2008,2
```

Using Time Information: Trends for drinking



Evaluation Methods

Intrinsic Evaluation

- The most common metric is to test their performance on word similarity

Evaluation Methods

Intrinsic Evaluation

- The most common metric is to test their performance on word similarity
- In particular, by computing correlation between an algorithm's word similarity scores and word similarity ratings assigned by humans.

Evaluation Methods

Intrinsic Evaluation

- The most common metric is to test their performance on word similarity
- In particular, by computing correlation between an algorithm's word similarity scores and word similarity ratings assigned by humans.
- Example benchmarks: Simlex-999, WordSim-353 etc.
- Other benchmarks: Selectional preferences, analogy testing etc.

Benchmarks

Rubenstein & Goodenough

- 65 noun pairs rated by 51 subjects on a 0-4 similarity scale and averaged
- E.g., *car-automobile 3.9; food-fruit 2.7; cord-smile 0.0*

Benchmarks

Rubenstein & Goodenough

- 65 noun pairs rated by 51 subjects on a 0-4 similarity scale and averaged
- E.g., *car-automobile 3.9; food-fruit 2.7; cord-smile 0.0*

WordSim-353

- 353 noun pairs, with ratings from 0 to 10 as given by humans; e.g. *(plane, car)* had an average rating of *5.77*.

Semantic Priming

- Hearing/reading a “related” prime facilitates access to a target in various lexical tasks
- You recognize/access the word *pear* faster if you heard/read *apple*
- Hodgson found similar amounts of priming for different semantic relations between primes and targets (23 pairs per relation):
  - synonyms (synonym): to dread/to fear
  - antonyms (antonym): short/tall
  - coordinates (coord): train/truck
  - super- and subordinate pairs (supersub): container/bottle
  - free association pairs (freeass): dove/peace
  - phrasal associates (phrasacc): vacant/building

Simulating semantic priming

For each related prime-target pair:

- measure cosine-based similarity between pair elements (e.g., to dread/to fear)
- take average of cosine-based similarity of target with other primes from same relation data-set (e.g., to value/to fear) as measure of similarity of target with unrelated items
- Similarity between related items should be significantly higher than average similarity between unrelated items

Other Evaluation benchmarks

Selectional Preferences

eat	villager	obj	1.7
eat	pizza	obj	6.8

Analogy

syntactic analogy		semantic analogy	
work	speak	brother	grandson
works	speaks	sister	granddaughter

$$\overrightarrow{\text{speaks}} \approx \overrightarrow{\text{works}} - \overrightarrow{\text{work}} + \overrightarrow{\text{speak}}$$

Two properties of representations in DSMs

- Distributed - Meaning is not represented in terms of some conceptual or formal symbols, but in terms of a multi-dimensional vector.
  - Vector dimensions are typically contexts
  - Semantic properties derive from global vector comparison (measuring their distance in space)

Two properties of representations in DSMs

- Distributed - Meaning is not represented in terms of some conceptual or formal symbols, but in terms of a multi-dimensional vector.
  - Vector dimensions are typically contexts
  - Semantic properties derive from global vector comparison (measuring their distance in space)
- Quantitative and gradual - Words differ not only for the contexts in which they appear, but also for the salience of these contexts.

Next Up

Word Vectors – *have taken over the distributional semantic models since their introduction.*