From distributional semantics to formal grammar and back

Chung-chieh Shan
Indiana University
Formal Grammar
11 August 2013
Approaches to semantics

“In order to say what a meaning *is*,
we may first ask what a meaning *does,*
and then find something that does that.”  —David Lewis
Approaches to semantics

“In order to say what a meaning is, we may first ask what a meaning does, and then find something that does that.” —David Lewis

Truth, entailment

Every person cried. ⊨ Every professor cried.
A person cried. ⊭ A professor cried.

Formal semantics

\[ \forall x. P x \rightarrow C x \]

\[ \lambda g. \forall x. P x \rightarrow g x \quad C \]

\[ \lambda f. \lambda g. \forall x. f x \rightarrow g x \quad P \]
Approaches to semantics

“In order to say what a meaning is, we may first ask what a meaning does, and then find something that does that.” —David Lewis

Concepts, similarity

ambulance $\sim$ battleship

ambulance $\sim$ bookstore

Distributional semantics

\[
\begin{array}{c|cccc}
\text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \text{accept} \\
\hline
\text{ambulance} & 27 & 10 & 50 & 17 & 130 & \ldots \\
\text{battleship} & 35 & 0 & 32 & 1 & 25 & \ldots \\
\text{bookstore} & 5 & 0 & 6 & 33 & 13 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\end{array}
\]
Approaches to semantics

“In order to say what a meaning is, we may first ask what a meaning does, and then find something that does that.” —David Lewis

What is meaning?

- Over-constrained problem?
  Meanings do lots of things.

- Under-constrained problem?
  Passive observations alone can’t distinguish what meanings are from how meanings are used.

- Seek unity in diversity . . .
  At least, meanings should support both entailment and similarity judgments, possibly with the help of world knowledge.
Distributional semantics for entailment among words

For each word $w$, rank contexts $c$ by descending $\frac{\Pr(c \mid w)}{\Pr(c)} > 1$.

“pointwise mutual information”
Distributional semantics for entailment among words

For each word $w$, rank contexts $c$ by descending $\frac{\Pr(c \mid w)}{\Pr(c)} > 1$.

“pointwise mutual information”

parent

argcount$_n$ arglist$_n$ arglist$_j$ phane$_n$ specity$_n$ qdisc$_n$ carthy$_n$
parents-to-be$_n$ non-resident$_j$ step-parent$_n$ tc$_n$ ballons$_n$
eliza$_n$ symptoms$_n$ adoptive$_j$ stepparent$_n$ nonresident$_j$
home-school$_n$ scabrid$_n$ petiolule$_n$ ...

person

anglia$_n$ first-mentioned$_j$ unascertained$_j$ enure$_v$
deposit-taking$_j$ bonis$_n$ iconclass$_j$ cotswolds$_n$ aforesaid$_n$
haver$_v$ foresaid$_j$ gha$_n$ sub-paragraphs$_n$ enacted$_j$ geest$_j$
non-medicinal$_j$ sub-paragraph$_n$ intimation$_n$ arrestment$_n$
incumbrance$_n$ ...

professor

william$_n$ extraordinarius$_n$ ordinarius$_n$ francis$_n$ reid$_n$
emeritus$_n$ emeritus$_j$ derwent$_n$ regius$_n$ laurence$_n$ edward$_n$
carisoprodol$_n$ adjunct$_j$ winston$_n$ privatdozent$_j$ edward$_j$
xanax$_n$ tenure$_v$ cialis$_n$ florence$_n$ ...

Distributional semantics for entailment among words

Context overlap with word 2

Context rank of word 1

- parent-person
- professor-person
- professor-parent
- person-parent
- person-professor
- parent-professor

More sophisticated: Kullback-Leibler divergence, skew divergence (Lee), \( \text{balAPinc} \) (Kotlerman et al.), . . .
Distributional semantics for entailment among words

More sophisticated: Kullback-Leibler divergence, skew divergence (Lee), balAPinc (Kotlerman et al.), ...
Sparse data strikes back

Successes for words and short phrases:
- similarity
- entailment
- sentiment

‘common sense’ from noisy large corpora

For long, rare, episodic phrases and sentences, need
- syntactic structure
- semantic reference
- pragmatic context
- grounding in other information sources

‘linguistic generalization’ from poor stimulus

This need goes way back—
Information retrieval started with bag of terms in each document. Stopwords, stemming, tagging; TF-IDF.

<table>
<thead>
<tr>
<th></th>
<th>abandon</th>
<th>abdominal</th>
<th>ability</th>
<th>academic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>10</td>
<td>50</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Information retrieval started with bag of terms in each document. Stopwords, stemming, tagging; TF-IDF. Dimensionality reduction reveals topics.

From documents $\times$ terms to words $\times$ contexts

$$\text{abandon} \quad \text{abdominal} \quad \text{ability} \quad \text{academic} \quad \ldots$$

$$\begin{pmatrix}
27 & 10 & 50 & 17 & \ldots \\
35 & 0 & 32 & 1 & \ldots \\
5 & 0 & 6 & 33 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}$$

$$=$$

$$\begin{pmatrix}
\vdots
\end{pmatrix} \times \begin{pmatrix}
\text{abandon} \\
\text{abdominal} \\
\text{ability} \\
\text{academic} \\
\ldots
\end{pmatrix}$$
From documents $\times$ terms to words $\times$ contexts

Information retrieval started with bag of terms in each document. Stopwords, stemming, tagging; TF-IDF. Dimensionality reduction reveals topics. Now rows are phrases and columns are contexts.

\[
\begin{bmatrix}
\text{ambulance} & 27 & 10 & 50 & 17 & \ldots \\
\text{battleship} & 35 & 0 & 32 & 1 & \ldots \\
\text{bookstore} & 5 & 0 & 6 & 33 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
\text{abandon} \\
\text{abdominal} \\
\text{ability} \\
\text{academic} \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
\text{ambulance} \\
\text{battleship} \\
\text{bookstore} \\
\vdots \\
\end{bmatrix}
\times
\begin{bmatrix}
\text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \ldots \\
\end{bmatrix}
\]
Composite phrases?

Need structure: substitution? locality? types? compositionality?

however, individuals with autism also have abnormal brain activation in many circuits outside the MNS and the MNS theory does not explain the normal performance of autistic children on imitation tasks that involve a goal or object.

To cope with sparse data, NLP (parsing, translation, compression) applies linguistic insight (typing, factoring, smoothing).
Composite phrases?

Need structure: substitution? locality? types? compositionality?

However, individual with autism also have abnormal brain activation in many circuits outside the MNS and the MNS theory does not explain the normal performance of autistic children on imitation tasks that involve a goal or object.
Composite phrases?

Need structure: substitution? locality? types? compositionality?

However, individuals with autism also have abnormal brain activation in many circuits outside the MNS and the MNS theory does not explain the normal performance of autistic children on imitation tasks that involve a goal or object.

To cope with sparse data, NLP (parsing, translation, compression) applies linguistic insight (typing, factoring, smoothing).
phrasal entailment

distributional semantics

language models

formal grammar
Above the word level

Phrases have corpus distributions too!

- **N** cat
- **AN** white cat
- **QN** every cat
Above the word level

Phrases have corpus distributions too!  But $N \approx AN \not\approx QN$

<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>N</th>
<th>AN</th>
<th>QN</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>cat</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>AN</td>
<td>white cat</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>QN</td>
<td>every cat</td>
<td>QP</td>
<td></td>
</tr>
</tbody>
</table>
Above the word level

Phrases have corpus distributions too!  But $\text{N} \approx \text{AN} \not\approx \text{QN}$

<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>Semantic type</th>
</tr>
</thead>
<tbody>
<tr>
<td>N  cat</td>
<td>N ( e \rightarrow t )</td>
</tr>
<tr>
<td>AN white cat</td>
<td>N ( e \rightarrow t )</td>
</tr>
<tr>
<td>QN every cat</td>
<td>QP ((e \rightarrow t) \rightarrow t)</td>
</tr>
</tbody>
</table>
Above the word level

<table>
<thead>
<tr>
<th></th>
<th>Syntactic category</th>
<th>Semantic type</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>cat</td>
<td>N</td>
</tr>
<tr>
<td>AN</td>
<td>white cat</td>
<td>N</td>
</tr>
<tr>
<td>AAN</td>
<td>big white cat</td>
<td>N</td>
</tr>
<tr>
<td>QN</td>
<td>every cat</td>
<td>QP</td>
</tr>
<tr>
<td>QAN</td>
<td>every big cat</td>
<td>QP</td>
</tr>
<tr>
<td>* AQN</td>
<td>big every cat</td>
<td>(e → t) → t</td>
</tr>
<tr>
<td>* QQN</td>
<td>some every cat</td>
<td>(e → t) → t</td>
</tr>
</tbody>
</table>

Phrases have corpus distributions too! But $N \approx AN \nleq QN$
Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among logical words rather than content words?

Different entailment relations at different semantic types?

\[\text{train} \rightarrow \text{test}\]

\[\begin{align*}
\text{AN} \models \text{N} & \quad \text{big cat} \rightarrow \text{cat} \\
\text{N} \models \text{N} & \quad \text{dog} \rightarrow \text{animal}
\end{align*}\]
Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among **logical words** rather than content words?

Different entailment relations at different semantic types?

\[ \text{AN} \models \text{N} \quad \text{big cat} \quad \text{cat} \quad \text{N} \models \text{N} \quad \text{dog} \quad \text{animal} \]

\[ \text{QN} \models \text{QN} \quad \text{many dogs} \quad \text{some dogs} \quad \text{QN} \models \text{QN} \quad \text{all cats} \quad \text{several cats} \]
Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among **logical words** rather than content words?

Different entailment relations at different **semantic types**?

\[ \text{AN} \models \text{N} \]
\[ \text{big cat} \Rightarrow \text{cat} \]

\[ \text{N} \models \text{N} \]
\[ \text{dog} \Rightarrow \text{animal} \]

\[ \text{QN} \models \text{QN} \]
\[ \text{many dogs} \Rightarrow \text{some dogs} \]

\[ \text{QN} \models \text{QN} \]
\[ \text{all cats} \Rightarrow \text{several cats} \]
Our semantic space

BNC, WackyPedia, ukWaC

\[ \text{TreeTagger (Schmid)} \]

lemmatized, POS-tagged tokens (2.8G)

\[ \text{words and phrases in the same sentence} \]

most frequent

A, N, V (27K)

\[ \begin{pmatrix}
\text{AN} \\
\text{QN} \\
\text{A} \\
\text{Q} \\
\text{N} \\
\end{pmatrix}
\]

\[ \#(c, w) \]
Our semantic space

BNC, WackyPedia, ukWaC

\[ \downarrow \text{TreeTagger (Schmid)} \]

lemmatized, POS-tagged tokens (2.8G)

\[ \downarrow \text{words and phrases in the same sentence} \]

most frequent

A, N, V (27K)

\[ \begin{pmatrix}
\text{AN} \\
\text{QN} \\
\text{AQ} \\
\text{N} \\
(48K)
\end{pmatrix}
\]

\#(c, w)

PMI

\[ \log \frac{\Pr(c \mid w)}{\Pr(c)} \]

SVD

\[ U\tilde{\Sigma} \]

(300)
Our semantic space

BNC, WackyPedia, ukWaC

\[ \text{TreeTagger (Schmid)} \]

lemmatized, POS-tagged tokens (2.8G)

\[ \text{words and phrases in the same sentence} \]

most frequent

A, N, V (27K)

\[ \text{AN} \]

\[ \text{QN} \]

\[ \text{A} \]

\[ \text{Q} \]

\[ \text{N} \]

(48K)

\[ \#(c, w) \]

\[ \log \frac{\Pr(c|w)}{\Pr(c)} \]

\[ \text{balAPinc} \]

\[ \text{SVM} \]

(300)
Our entailment classifiers

\[
\text{PMI} \rightarrow \left( \log \frac{\Pr(c|w)}{\Pr(c)} \right)
\]
Our entailment classifiers

\[
\log \frac{\Pr(c|w)}{\Pr(c)}
\]
Our entailment classifiers

\[ \log \frac{\Pr(c|w)}{\Pr(c)} \]

PMI

Train Test

AN \models N

QN \models QN

SVD

\[ ? \subseteq \]

\[ \text{(Kotlerman et al.)} \rangle \text{threshold?} \]
Our entailment classifiers

\[
\log \frac{\Pr(c \mid w)}{\Pr(c)}
\]

PMI

\[
\text{balAPinc}
\]

(Kotlerman et al.)
Our entailment classifiers

\[
\log \frac{\Pr(c|w)}{\Pr(c)}
\]

PMI

\[0 \leq \text{balAPinc} \leq 1\]

> threshold?
Our entailment classifiers

\[ \log \frac{\Pr(c \mid w)}{\Pr(c)} \]

PMI

Train | Test
-----|-----
AN ⊨ N | N ⊨ N
QN ⊨ QN | QN ⊨ QN
AN ⊨ N | QN ⊨ QN

\[ 0 \leq \text{balAPinc} \leq 1 \]

> threshold?
Our entailment classifiers

\[
\log \frac{\Pr(c|w)}{\Pr(c)}
\]

PMI

\[0 \leq \text{balAPinc} \leq 1\]

> threshold?

SVD

\[U\tilde{\Sigma}\]

SVM
(cubic)

outperformed naïve Bayes, \(k\)NN
Our data sets

WordNet

↓

pope ⊨ spiritual_leader
spiritual_leader ⊨ leader

cat ⊨ feline
feline ⊨ carnivore

: 
Our data sets

WordNet

\[
pope \models \text{leader}
\]
\[
cat \models \text{carnivore}
\]
\[
\vdash (1385)
\]
Our data sets

WordNet

pope ≡ leader

cat ≡ carnivore

leader ⊭ pope

cat ⊭ leader

invert

resample

N = N

E = E

Q = Q

N = N

(e ! t)

train

test

train

test
Our data sets

most frequent

↓

big
former

 đèn (300)

WordNet

↓

pope ⊨ leader

\[ \text{cat} \equiv \text{carnivore} \]

\[ \vdash (1385) \]

invert

resample

leader ⊬ pope
\[ \vdash (1385) \]

cat ⊬ leader

\[ \vdash (1385) \]
Our data sets

most frequent
  ↓
  big
  former
  ⊨ (256)

WordNet
  ↓
  pope ⊨ leader
  cat ⊨ carnivore
  ⊨ (1385)

invert

leader ⊭ pope
  cat ⊭ leader
  ⊨ (1385)

resample

all
  both
  each
  either
  every
  few
  many
  most
  much
  no
  several
  some

big apple ⊨ apple
  big shirt ⊨ shirt
  ⊨ (1246)

leader ⊭ pope
  cat ⊭ leader
  ⊨ (1385)

big apple ⊭ shirt
  big shirt ⊭ apple
  ⊨ (1244)

pope
  leader
  cat
  carnivore
  ⊨ (6402)
Our data sets

- **most frequent**
  - big
  - former
    - big apple $\models$ apple
    - big shirt $\models$ shirt
      - big apple $\not\models$ shirt
      - big shirt $\not\models$ apple
    - big apple $\not\models$ shirt
    - big shirt $\not\models$ apple
  - (256)

- **BLESS**
  - apple
  - shirt
    - (200)

- **WordNet**
  - pope $\models$ leader
  - cat $\models$ carnivore
    - leader $\not\models$ pope
    - cat $\not\models$ leader
    - pope $\not\models$ leader
    - cat $\not\models$ leader
  - (1385)
Our data sets

most frequent

**big**

former

\(\vdash (256)\)

BLESS

**apple**

**shirt**

\(\vdash (200)\)

WordNet

**pope** \(\models\) **leader**

**cat** \(\models\) **carnivore**

\(\vdash (1385)\)

most frequent

**all**

**both**

**each**

**every**

**few**

**many**

**most**

**much**

**no**

**several**

**some**

\(\vdash:\)
Our data sets

most frequent

big former

apple shirt

\( \downarrow \)

BLESS

apple shirt

\( \downarrow \)

WordNet

pope ⊨ leader
cat ⊨ carnivore

\( \downarrow \)

most frequent

all ⊨ some
many ⊨ several

\( \downarrow \)

some ⊭ every
both ⊭ many

\( \downarrow \)

big apple ⊭ shirt
big shirt ⊭ apple

\( \downarrow \)

(256)
(200)
(1246)
(1244)
(1385)
(1385)
(13)
(17)
(6402)
(7537)
(8455)
Our data sets

most frequent

<table>
<thead>
<tr>
<th>BLESS</th>
<th>WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>big former</td>
<td>pope ⊨ leader</td>
</tr>
<tr>
<td>apple shirt</td>
<td>cat ⊨ carnivore</td>
</tr>
<tr>
<td>big apple ⊨ apple</td>
<td>leader ⊭ pope</td>
</tr>
<tr>
<td>big shirt ⊨ shirt</td>
<td>cat ⊭ leader</td>
</tr>
</tbody>
</table>

most frequent

| all ⊨ some |
| many ⊨ several |
| some ⊭ every |
| both ⊭ many |

all cat ⊭ some cat
many cat ⊭ several cat
some cat ⊭ every cat
both cat ⊭ many cat

all cat ⊭ every leader
both cat ⊭ many leader
Our data sets

most frequent

- big
- former
  - (256)

BLESS

- apple
- shirt
  - (200)

WordNet

- pope ⊨ leader
- cat ⊨ carnivore
  - (1385)

- big apple ⊭ apple
- big shirt ⊭ shirt
  - (1246)

most frequent

- all ⊨ some
- many ⊨ several
  - (13)

- some ⊭ every
- both ⊭ many
  - (17)

- big apple ⊭ shirt
- big shirt ⊭ apple
  - (1244)

- invert
- resample
- resample

- pope ⊭ leader
- cat ⊭ leader
  - (1385)

- all cat ⊨ some cat
- many cat ⊨ several cat
  - (7537)

- some cat ⊭ every cat
- both cat ⊭ many cat
  - (8455)

- (e → t) → t
Our data sets
## Results at noun type

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Accuracy</th>
<th>(95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM&lt;sub&gt;upper&lt;/sub&gt;</td>
<td>88.6</td>
<td>88.6</td>
<td>88.5</td>
<td>88.6</td>
<td>(87.3–89.7)</td>
</tr>
<tr>
<td>balAPinc&lt;sub&gt;AN ⊨ N&lt;/sub&gt;</td>
<td>65.2</td>
<td>87.5</td>
<td>74.7</td>
<td>70.4</td>
<td>(68.7–72.1)</td>
</tr>
<tr>
<td>balAPinc&lt;sub&gt;upper&lt;/sub&gt;</td>
<td>64.4</td>
<td>90.0</td>
<td>75.1</td>
<td>70.1</td>
<td>(68.4–71.8)</td>
</tr>
<tr>
<td>SVM&lt;sub&gt;AN ⊨ N&lt;/sub&gt;</td>
<td>69.3</td>
<td>69.3</td>
<td>69.3</td>
<td>69.3</td>
<td>(67.6–71.0)</td>
</tr>
<tr>
<td>cos(N&lt;sub&gt;1&lt;/sub&gt;, N&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>57.7</td>
<td>57.6</td>
<td>57.5</td>
<td>57.6</td>
<td>(55.8–59.5)</td>
</tr>
<tr>
<td>fq(N&lt;sub&gt;1&lt;/sub&gt;) &lt; fq(N&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>52.1</td>
<td>52.1</td>
<td>51.8</td>
<td>53.3</td>
<td>(51.4–55.2)</td>
</tr>
</tbody>
</table>
### Holding out QN data

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>both</th>
<th>each</th>
<th>either</th>
<th>every</th>
<th>few</th>
<th>many</th>
<th>most</th>
<th>much</th>
<th>no</th>
<th>several</th>
<th>some</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>all</strong></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>both</strong></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>each</strong></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>either</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>every</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>few</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>many</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>most</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>much</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>no</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>several</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>some</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Holding out QN data

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>both</th>
<th>each</th>
<th>either</th>
<th>every</th>
<th>few</th>
<th>many</th>
<th>most</th>
<th>much</th>
<th>no</th>
<th>several</th>
<th>some</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>both</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>each</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>either</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>every</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>few</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>many</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>most</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>much</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>several</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>some</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

.pair-out
# Holding out QN data

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>both</th>
<th>each</th>
<th>either</th>
<th>every</th>
<th>few</th>
<th>many</th>
<th>most</th>
<th>much</th>
<th>no</th>
<th>several</th>
<th>some</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>both</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>each</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>either</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>every</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>few</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>many</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>most</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>much</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>no</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>several</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>some</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
# Results at quantifier type

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Accuracy (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM\textsuperscript{pair-out}</td>
<td>76.7</td>
<td>77.0</td>
<td>76.8</td>
<td>78.1 (77.5–78.8)</td>
</tr>
<tr>
<td>SVM\textsuperscript{quantifier-out}</td>
<td>70.1</td>
<td>65.3</td>
<td>68.0</td>
<td>71.0 (70.3–71.7)</td>
</tr>
<tr>
<td>SVM\textsuperscript{Q}\textsubscript{pair-out}</td>
<td>67.9</td>
<td>69.8</td>
<td>68.9</td>
<td>70.2 (69.5–70.9)</td>
</tr>
<tr>
<td>SVM\textsuperscript{Q}\textsubscript{quantifier-out}</td>
<td>53.3</td>
<td>52.9</td>
<td>53.1</td>
<td>56.0 (55.2–56.8)</td>
</tr>
<tr>
<td>(\text{cos}(QN_1, QN_2))</td>
<td>52.9</td>
<td>52.3</td>
<td>52.3</td>
<td>53.1 (52.3–53.9)</td>
</tr>
<tr>
<td>balAP\textsubscript{inc}_AN \vDash N</td>
<td>46.7</td>
<td>5.6</td>
<td>10.0</td>
<td>52.5 (51.7–53.3)</td>
</tr>
<tr>
<td>SVM\textsubscript{AN \vDash N}</td>
<td>2.8</td>
<td>42.9</td>
<td>5.2</td>
<td>52.4 (51.7–53.2)</td>
</tr>
<tr>
<td>(\text{fq}(QN_1) &lt; \text{fq}(QN_2))</td>
<td>51.0</td>
<td>47.4</td>
<td>49.1</td>
<td>50.2 (49.4–51.0)</td>
</tr>
<tr>
<td>balAP\textsubscript{inc}\textsubscript{upper}</td>
<td>47.1</td>
<td>100</td>
<td>64.1</td>
<td>47.2 (46.4–47.9)</td>
</tr>
</tbody>
</table>
## Holding out each quantifier

<table>
<thead>
<tr>
<th>Quantifier</th>
<th>Instances</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \models )</td>
<td>( \not\models )</td>
</tr>
<tr>
<td>each</td>
<td>656</td>
<td>656</td>
</tr>
<tr>
<td>every</td>
<td>460</td>
<td>1322</td>
</tr>
<tr>
<td>much</td>
<td>248</td>
<td>0</td>
</tr>
<tr>
<td>all</td>
<td>2949</td>
<td>2641</td>
</tr>
<tr>
<td>several</td>
<td>1731</td>
<td>1509</td>
</tr>
<tr>
<td>many</td>
<td>3341</td>
<td>4163</td>
</tr>
<tr>
<td>few</td>
<td>0</td>
<td>461</td>
</tr>
<tr>
<td>most</td>
<td>928</td>
<td>832</td>
</tr>
<tr>
<td>some</td>
<td>4062</td>
<td>3145</td>
</tr>
<tr>
<td>no</td>
<td>0</td>
<td>714</td>
</tr>
<tr>
<td>both</td>
<td>636</td>
<td>1404</td>
</tr>
<tr>
<td>either</td>
<td>63</td>
<td>63</td>
</tr>
</tbody>
</table>

| Total      | 15074     | 16910   | 9849      | 12870     | (71%) |
Interim summary

Entailment among composite phrases rather than nouns.  
(Cheap training data!)

Entailment among logical words rather than content words.  
(Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types.  
(Prediction from formal semantics.)

\[ \text{AN} \models \text{N} \quad \longrightarrow \quad \text{N} \models \text{N} \]
\[ \text{big cat} \quad \text{cat} \quad \longrightarrow \quad \text{dog} \quad \text{animal} \]

\[ \text{QN} \models \text{QN} \quad \longrightarrow \quad \text{QN} \models \text{QN} \]
\[ \text{many dogs} \quad \text{some dogs} \quad \longrightarrow \quad \text{all cats} \quad \text{several cats} \]
Interim summary

Entailment among **composite phrases** rather than nouns.  
(Cheap training data!)

Entailment among logical words rather than content words.  
(Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types.  
(Prediction from formal semantics.)

![Diagram showing entailment relations between groups of words](image-url)
Interim summary

Entailment among **composite phrases** rather than nouns. (Cheap training data!)

Entailment among **logical words** rather than content words. (Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types. (Prediction from formal semantics.)

\[ \text{AN} \models \text{N} \quad \text{big cat} \quad \text{cat} \]
\[ \text{N} \models \text{N} \quad \text{dog} \quad \text{animal} \]
\[ \times \]
\[ \text{QN} \models \text{QN} \quad \text{many dogs} \quad \text{some dogs} \]
\[ \text{QN} \models \text{QN} \quad \text{all cats} \quad \text{several cats} \]
Interim summary

Entailment among **composite phrases** rather than nouns.
(Cheap training data!)

Entailment among **logical words** rather than content words.
(Part of Recognizing Textual Entailment?)

Different entailment relations at different **semantic types**.
(Prediction from formal semantics.)
Interim summary

Entailment among **composite phrases** rather than nouns. (Cheap training data!) ❍ Practical import

Entailment among **logical words** rather than content words. (Part of Recognizing Textual Entailment?) ❍ Practical import

Different entailment relations at different **semantic types**. (Prediction from formal semantics.)

\[
\begin{align*}
AN & \models_{N} N \\
\text{big cat} & \quad \text{cat}
\end{align*}
\]

\[
\begin{align*}
N & \models_{N} N \\
\text{dog} & \quad \text{animal}
\end{align*}
\]

\[
\begin{align*}
QN & \models_{QN} QN \\
\text{many dogs} & \quad \text{some dogs}
\end{align*}
\]

\[
\begin{align*}
QN & \models_{QN} QN \\
\text{all cats} & \quad \text{several cats}
\end{align*}
\]
Interim summary

Entailment among **composite phrases** rather than nouns. (Cheap training data!) 

Entailment among **logical words** rather than content words. (Part of Recognizing Textual Entailment?)  

Different entailment relations at different **semantic types**. (Prediction from formal semantics.)

Ongoing work:
- How does the SVM work?
- Missing experiments?
- How to compose semantic vectors?
A language model is a virtual infinite corpus: not frequencies observed but probabilities estimated.

Let the distributional meaning of a phrase $w$ be the probability distribution over its contexts $c$.

$$[w] = \lambda c. \frac{\Pr(c[w])}{\sum_{c'} \Pr(c'[w])}$$

$$[\text{red army}] = \lambda(l, r). \frac{\Pr(l \text{ red army } r)}{\sum_{(l', r')} \Pr(l' \text{ red army } r')}$$

$$[\text{red } w] = \lambda(l, r). \frac{[w](l \text{ red }, r)}{\sum_{(l', r')} [w](l' \text{ red }, r')}$$

Probabilities from any model: bag of words, Markov, PCFG. . . Pass the buck. Language models and corpora can (should?) include world reference in utterance context.
A language model is a virtual infinite corpus: not frequencies observed but probabilities estimated.

Let the distributional meaning of a phrase \( w \) be the probability distribution over its contexts \( c \).

\[
[w] = \lambda(c) \frac{\Pr(c[w])}{\sum_{c'} \Pr(c'[w])}
\]

\[
[\text{red army}] = \lambda(l, r) \frac{\Pr(l \text{ red army } r)}{\sum_{(l', r')} \Pr(l' \text{ red army } r)}
\]

\[
[\text{red } w] = \lambda(l, r) \frac{[w](l \text{ red, } r)}{\sum_{(l', r')} [w](l' \text{ red, } r')}
\]

Probabilities from any model: bag of words, Markov, PCFG… Pass the buck. Language models and corpora can (should?) include world reference in utterance context.
From Penn Treebank to distributional semantics

Penn Treebank

Collins model

PCFG

\[ \cap \{l \text{ phone } r\} \]

\[ \cap \{l \text{ radio } r\} \]

PCFG

PCFG

semantic relations, KL divergence, etc.
Intersection grammars reveal meaning?

Random sentences


have/VBP [lately/RB [[in/IN [July/NNP [[[weighted/JJ large/JJ] exchange/NN] for/IN] clients/NNS]] been/VBN]]

was/VBD [[/, [Mr./NNP Bush/NNP]] said/VBD]

[plunged/VBD [4/CD ,/,]]

[[start/NN of/IN] [was/VBD [me/PRP [1.9/CD pence/NN]]]]

has/VBZ [[/, they/PRP] [see/VBP [aided/VBN [[in/IN [[some/DT seconds/NNS] [[of/IN] executive/NN [[[a/DT share/NN] [[yesterday/NN [[the/DT last/JJ] market/NN]]
 acted/VBD]] [[the/DT advance/NN] revenue/NN]]]] [a/DT share/NN]]]] [by/IN [[[[the/DT pound/NN] and/CC] Exchange/NNP]]]]]]

would/MD

[[Dec./NNP 28/CD] [announced/VBD that/IN]]

Intersection grammars reveal meaning?

Random sentences containing “car dealer”

```
[[many/JJ [[the/DT car/NN] dealer/NN] Developers/NNS]] ,[,] are/VBP


[resigned/VBD [to/TO] [price/VB] ((([[the/DT car/NN] dealer/NN] [from/IN] [Mr./NNP] +unknown+/NNP] on/IN]] [through/RP] [such/PDT a/DT] offering/NN]])] [showing/VBG ((([[The/DT junk/NN] defense/NN] measure/NN] [bank/NN known/VBN]]))]]

((((((((still/RB [the/DT Gardens/NNPS] life/NN] [initial/JJ transaction/NN]] [[a/DT few/JJ] arrangement/NN] administration/NN] [The/DT group/NN]] [[the/DT first/NN] price/NN] [of/IN] ,[,] [is/VBZ [[car/NN] dealer/NN] of/IN] [[$/$ +unknown+/CD] [the/DT futures/NNS]]))] was/VBD [going/VBG [[against/IN
```
Intersection grammars reveal meaning?

Random sentences containing “drug dealer”

In past five structural +unknown NN of Allied stock [+unknown +unknown farmer +unknown NN] [the most JJS who [drag require because IN] because of the DT company [a year [+unknown +unknown +unknown cooperative children]]] [this DT +unknown +unknown OTC market for IN] [The DT company [The DT drug dealer] soared reducing VBG , , Monday NN]

[[rose VBD from IN [[$/ [ +unknown CD million ] million ] +unknown +CD million billion 15.6 CD ]]]

[a DT share NN , , today NN tickets NN ] [[ AD deep JJ series of IN fact [last JJ car [that WDT] with IN looks NN ] [seven CD cents NN [a DT share NN ]]]] [on IN [[The DT +unknown +unknown ownership NN or CC yesterday NN]] [[another DT year [NN] price NN above IN — : ]] [was VBD [against IN him PRP although IN]] [but CC [[the DT following JJ week NN] was VBD [a DT specific JJ +unknown +unknown short-term JJ] Treasury NN economic JJ trade NN [the DT FDA 1/ NNP ] . ]] [marks NNS [[the DT coming VBG early JJ next JJ year NN] [little RB] , , ] [and CC and CC , , CC ] [+unknown +NNP +unknown NNS] has VBZ n’t RB] received VBD [a DT serious JJ drag NN on IN] [now RB [when WRB [[the DT drug NN dealer NN] pay VB] [[[Sun NNP Jeep Stoll NNP] entered VBD [MCI NNP +unknown - NN] U.S. NNP] hurt VBP] [[the DT [...]]] 22/33
Intersection grammars reveal meaning?

Random sentences containing “card dealer”


Intersection grammars reveal meaning?

Top sentences

2e−2  [said/VBD]
1e−2  [is/VBZ]
8e−3  [was/VBD]
7e−3  [are/VBP]
6e−3  [has/VBZ]
    ...
5e−4  [[[he/PRP]] [said/VBD]]
    ...
4e−4  [[[he/PRP]] [says/VBZ]]
    ...
9e−5  [[[the/DT] company/NN]] [said/VBD]]
    ...
7e−5  [[[she/PRP]] [says/VBZ]]
    ...
2e−6  [[[Mr./NNP] Inc./NNP]] [said/VBD]]
2e−6  [[is/VBZ [first/JJ]]]
2e−6  [more/RBR]
2e−6  [has/VBZ [failed/VBN]]]
Intersection grammars reveal meaning?

Top sentences containing “car dealer”

2e−10 [[[[car/NN] dealer/NN]]] [said/VBD]
8e−11 [[[[a/DT] car/NN] dealer/NN]]] [said/VBD]
7e−11 [[[[the/DT] car/NN] dealer/NN]]] [said/VBD]
5e−11 [[[[the/DT] car/NN] dealer/NN]]] [said/VBD]
4e−11 [[car/NN] dealer/NN]]
3e−11 [[car/NN] dealer/NN]]] [said/VBD]
3e−11 [[says/VBZ] [[[[car/NN] dealer/NN]]]]
3e−11 [[[[car/NN] dealer/NN]]] ,/,[ said/VBD]
3e−11 [[,] [[[car/NN] dealer/NN]]] [said/VBD]
2e−11 [[[[a/DT] car/NN] dealer/NN]]] [said/VBD]
2e−11 [[[[car/NN] dealer/NN]]] [was/VBD]
2e−11 [[[a/DT] car/NN] dealer/NN]]
2e−11 [[[the/DT] car/NN] dealer/NN]]
2e−11 [[[[a/DT] car/NN] dealer/NN]]] [said/VBD]
2e−11 [[[[the/DT] car/NN] dealer/NN]]] [said/VBD]
2e−11 [[says/VBZ] [[[[a/DT] car/NN] dealer/NN]]]]
1e−11 [[says/VBZ] [[[[the/DT] car/NN] dealer/NN]]]]
1e−11 [[[[car/NN] dealer/NN]]] [says/VBZ]
1e−11 [[[[the/DT] car/NN] dealer/NN]]
1e−11 [[[[a/DT] car/NN] dealer/NN]]] ,/,[ said/VBD]
1e−11 [[,] [[[a/DT] car/NN] dealer/NN]]] [said/VBD]
Intersection grammars reveal meaning?

Top sentences containing “drug dealer”

6e—9 [[[drug/NN dealer/NN]]] [said/VBD]
6e—9 [[[the/DT drug/NN dealer/NN]]] [said/VBD]
2e—9 [[[the/DT drug/NN dealer/NN]]] [said/VBD]
2e—9 [[[drug/NN dealer/NN]]]
2e—9 [[[the/DT drug/NN dealer/NN]]]
1e—9 [[[The/DT drug/NN dealer/NN]]] [said/VBD]
1e—9 [[[says/VBZ] [[[drug/NN dealer/NN]]]]]
1e—9 [[[says/VBZ] [[[the/DT drug/NN dealer/NN]]]]]
1e—9 [[[a/DT drug/NN dealer/NN]]] [said/VBD]
1e—9 [[[drug/NN dealer/NN]]] , [said/VBD]
1e—9 [[[drug/NN dealer/NN]]] [said/VBD]
1e—9 [[[the/DT drug/NN dealer/NN]]] , [said/VBD]
8e—10 [[[drug/NN dealer/NN]]] [was/VBD]
8e—10 [[[the/DT drug/NN dealer/NN]]] [was/VBD]
7e—10 [[[a/DT drug/NN dealer/NN]]] [said/VBD]
5e—10 [[[the/DT drug/NN dealer/NN]]]
5e—10 [[[drug/NN dealer/NN]]] [says/VBZ]
5e—10 [[[the/DT drug/NN dealer/NN]]] [says/VBZ]
4e—10 [[[drug/NN dealer/NN]]] [said/VBD]
4e—10 [[[drug/NN dealer/NN]]] [noted/VBD]
Intersection grammars reveal meaning?

Top sentences containing “card dealer”

2e−11 [[[[the/DT] card/NN] dealer/NN]] [said/VBD]
1e−11 [[[[card/NN] dealer/NN]] [said/VBD]]
8e−12 [[[[a/DT] card/NN] dealer/NN]] [said/VBD]
6e−12 [[[[the/DT] card/NN] dealer/NN]] [said/VBD]
4e−12 [[[[the/DT] card/NN] dealer/NN]]
4e−12 [[[[a/DT] card/NN] dealer/NN]] [said/VBD]
3e−12 [[[card/NN] dealer/NN]]
3e−12 [[[says/VBZ] [[[the/DT] card/NN] dealer/NN]]]]
3e−12 [[[The/DT] card/NN] dealer/NN]] [said/VBD]]
3e−12 [[[[the/DT] card/NN] dealer/NN]] ,/[,] [said/VBD]]
3e−12 [[,,] [[the/DT] card/NN] dealer/NN]] [said/VBD]]
3e−12 [[[says/VBZ] [[[card/NN] dealer/NN]]]]
2e−12 [[[[the/DT] card/NN] dealer/NN]] [said/VBD]
2e−12 [[[a/DT] card/NN] dealer/NN]]
2e−12 [[[The/DT] card/NN] dealer/NN]] [said/VBD]]
2e−12 [[[card/NN] dealer/NN]] ,/[,] [said/VBD]
2e−12 [[,,] [[[card/NN] dealer/NN]]] [said/VBD]]
2e−12 [[[[the/DT] card/NN] dealer/NN]] [was/VBD]]
2e−12 [[[card/NN] dealer/NN]] [said/VBD]
2e−12 [[[the/DT] card/NN] dealer/NN]]
2e−12 [[[card/NN] dealer/NN]] [was/VBD]]
Kullback-Leibler divergence

\[ D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{1}{Q(x)} - \sum_x P(x) \log \frac{1}{P(x)} \]

cross entropy

Example

20 samples from \( P \):

\[ \text{encoded for } P: \quad \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \]

\[ \text{encoded for } Q: \quad \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \]
Kullback-Leibler divergence

\[ D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{1}{Q(x)} - \sum_x P(x) \log \frac{1}{P(x)} \]

cross entropy

entropy

Example

\[ P = \begin{array}{c} 0 \\ 10 \\ 110 \\ 111 \end{array} \quad Q = \begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \end{array} \]

20 samples from \( P \): \( \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \)

encoded for \( P \): 10001001101010011010101110110000111

encoded for \( Q \): 10000000001101000001100010110001000000011

KL divergence: 0.25 bits = 2.00 bits − 1.75 bits
From Penn Treebank to distributional semantics

Penn Treebank

Collins model

PCFG

\[ \cap \{l \text{ phone } r\} \]

\[ \cap \{l \text{ radio } r\} \]

PCFG

PCFG

semantic relations, KL divergence, etc.
From Penn Treebank to distributional semantics

$$\Pr(c, w)$$

- Bag-of-words model
  - $$\Pr(c)$$
  - $$\Pr(c)$$
  - KL divergence

- Collins NP model
  - $$\Pr(c)$$
  - PFSA
  - $$\Pr(c)$$
  - PFSA
  - KL divergence
Collins model

Lexicalized PCFG for parsing (1997)
Not for generation (Post & Gildea 2008)
Bikel (2004) exegesis

S(bought/VB)

NP(week/NN)  NP(Marks/NNP)  VP(bought/VB)

NPB(week/NN)  NPB(Marks/NNP)  bought/VB  NP(Brooks/NNP)

Last/JJ  week/NN  Marks/NNP  NPB(Brooks/NNP)

Brooks/NNP
Collins model

Lexicalized PCFG for parsing (1997)
Not for generation (Post & Gildea 2008)
Bikel (2004) exegesis

S(bought/VB)

NP(week/NN)  NP(Marks/NNP)  VP(bought/VB)

NPB(week/NN)  NPB(Marks/NNP)  bought/VB  NP(Brooks/NNP)

Last/JJ  week/NN  Marks/NNP  NPB(Brooks/NNP)

Brooks/NNP
Collins model

Lexicalized PCFG for parsing (1997)
Not for generation (Post & Gildea 2008)
Bikel (2004) exegesis

S(bought/VB)

NP(week/NN)  NP(Marks/NNP)  VP(bought/VB)

NPB(week/NN)  NPB(Marks/NNP)  bought/VB  NP(Brooks/NNP)

Last/JJ  week/NN  Marks/NNP  NPB(Brooks/NNP)

Brooks/NNP
Summary statistics

Standard English training set: Wall Street Journal §§02–21
- 39,832 sentences
- 950,028 word tokens
  - 44,113 unique words
  - 10,437 unique words that occur 6+ times
- 28 basic nonterminal labels
  - 42 parts of speech

Tiny for a corpus today.

Simplified Collins Model 1
- 575,936 nonterminals
  - 15,564 terminals
  - 12,611,676 rules

Big for a grammar today.
## Pilot evaluation using BLESS data set

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation</th>
<th>Relatum</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>coord</td>
<td>computer</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>radio</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>stereo</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>television</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>commodity</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>device</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>equipment</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>good</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>object</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>system</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>cable</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>dial</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>number</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>plastic</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>wire</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>choice</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>clearance</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>closing</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>entrepreneur</td>
</tr>
</tbody>
</table>

Baroni and Lenci Evaluation of Semantic Spaces (2011)

Only head nouns observed in corpus:

\[
\text{NP(phone/NN)} \quad \text{NPB(phone/NN)} \quad \text{phone/NN}
\]

Compute KL divergences among distributions over `modifier-nonterminal sequences`
Pilot evaluation using BLESS data set

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation</th>
<th>Relatum</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>coord</td>
<td>computer</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>radio</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>stereo</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>television</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>commodity</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>device</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>equipment</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>good</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>object</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>system</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>cable</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>dial</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>number</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>plastic</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>wire</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>choice</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>clearance</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>closing</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>entrepreneur</td>
</tr>
</tbody>
</table>

Baroni and Lenci Evaluation of Semantic Spaces (2011)

Only head nouns observed in corpus:

NP(phone/NN)  
|  
NPB(phone/NN)  
|  
phone/NN

Compute KL divergences among distributions over *modifier-nonterminal sequences*
Pilot evaluation using BLESS data set

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation</th>
<th>Relatum</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>coord</td>
<td>computer</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>radio</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>stereo</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>television</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>commodity</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>device</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>equipment</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>good</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>object</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>system</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>cable</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>dial</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>number</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>plastic</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>wire</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>choice</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>clearance</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>closing</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>entrepreneur</td>
</tr>
</tbody>
</table>

Baroni and Lenci Evaluation of Semantic Spaces (2011)

Only head nouns observed in corpus:

\[
\text{NP(phone/NN)} \rightarrow \text{NPB(phone/NN)} \rightarrow \text{phone/NN}
\]

Compute KL divergences among distributions over *modifier-nonterminal sequences*.
Pilot evaluation using BLESS data set

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relation</th>
<th>Relatum</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>173 coord</td>
<td>computer</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>radio</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>stereo</td>
</tr>
<tr>
<td>phone</td>
<td>coord</td>
<td>television</td>
</tr>
<tr>
<td>phone</td>
<td>125 hyper</td>
<td>commodity</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>device</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>equipment</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>good</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>object</td>
</tr>
<tr>
<td>phone</td>
<td>hyper</td>
<td>system</td>
</tr>
<tr>
<td>phone</td>
<td>490 mero</td>
<td>cable</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>dial</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>number</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>plastic</td>
</tr>
<tr>
<td>phone</td>
<td>mero</td>
<td>wire</td>
</tr>
<tr>
<td>phone</td>
<td>561 random-n</td>
<td>choice</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>clearance</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>closing</td>
</tr>
<tr>
<td>phone</td>
<td>random-n</td>
<td>entrepreneur</td>
</tr>
</tbody>
</table>

Baroni and Lenci Evaluation of Semantic Spaces (2011)

Only head nouns observed in corpus:

NP(phone/NN)

NPB(phone/NN) ← phone/NN

Compute KL divergences among distributions over modifier-nonterminal sequences
$D_{KL}(\text{Concept} \parallel \text{Relatum})$

$D_{KL}(\text{Relatum} \parallel \text{Concept})$

NP

\[ \to \]

NPB

\[ \to \]

NNS

$r_{np\_npb\_KL}$ by Relation (Kruskal–Wallis rank sum test $p=1.73002\times10^{-5}$)

Proportion $\leq x$

n:1288 m:61

Coord
Hyper
Mero
RandomN

$r_{np\_npb\_LK}$ by Relation (Kruskal–Wallis rank sum test $p=5.88196\times10^{-6}$)

Proportion $\leq x$

n:1288 m:61

Coord
Hyper
Mero
RandomN

$l_{np\_npb\_KL}$ by Relation (Kruskal–Wallis rank sum test $p=2.54141\times10^{-13}$)

Proportion $\leq x$

n:1340 m:9

Coord
Hyper
Mero
RandomN

$l_{np\_npb\_LK}$ by Relation (Kruskal–Wallis rank sum test $p=1.0453\times10^{-15}$)

Proportion $\leq x$

n:1340 m:9

Coord
Hyper
Mero
RandomN
$D_{KL}(\text{Concept} \ || \ Relatum)$

$D_{KL}(\text{Relatum} \ || \ Concept)$

Bag of words
Mann-Whitney-Wilcoxon rank sum test

Edges indicate $p < .01$

<table>
<thead>
<tr>
<th>$D_{KL}(\text{Concept} \mid \mid \text{Relatum})$</th>
<th>$D_{KL}(\text{Relatum} \mid \mid \text{Concept})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>coord hyper coord hyper</td>
</tr>
<tr>
<td>NPB</td>
<td>mero random-n mero random-n</td>
</tr>
<tr>
<td>NPB</td>
<td>coord hyper coord hyper</td>
</tr>
<tr>
<td>NPB</td>
<td>mero random-n mero random-n</td>
</tr>
<tr>
<td>NNS</td>
<td>coord hyper coord hyper</td>
</tr>
<tr>
<td>NPB</td>
<td>mero random-n mero random-n</td>
</tr>
</tbody>
</table>
Mann-Whitney-Wilcoxon rank sum test

Edges indicate $p < .01$

\[
D_{KL}(\text{Concept} \ || \ Relatum) \quad D_{KL}(\text{Relatum} \ || \ Concept)
\]

```
<table>
<thead>
<tr>
<th>coord</th>
<th>hyper</th>
</tr>
</thead>
<tbody>
<tr>
<td>coord</td>
<td>hyper</td>
</tr>
<tr>
<td>mero</td>
<td>random-n</td>
</tr>
<tr>
<td>mero</td>
<td>random-n</td>
</tr>
</tbody>
</table>
```

Bag of words
Summary

Distributional semantics from language models
  ▶ Estimate felicity *in context* from observed use
  ▶ Cope with sparse data using linguistic insight such as syntax
  ▶ Better distributional semantics from better language models?

Phrasal entailment from distributional semantics
  ▶ Logical words
  ▶ Semantic types

Thanks!
  ▶ Bolzano: Ngoc-Quynh Do, European Masters Program in Language and Communication Technologies
  ▶ Trento: Marco Baroni, Raffaella Bernardi, Roberto Zamparelli
  ▶ Rutgers: Jason Perry, Matthew Stone
  ▶ Cornell: John Hale, Mats Rooth
## Holding out each quantifier pair

<table>
<thead>
<tr>
<th>Quantifier pair</th>
<th>Instances</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>all = some</td>
<td>1054</td>
<td>1044 (99%)</td>
</tr>
<tr>
<td>all = several</td>
<td>557</td>
<td>550 (99%)</td>
</tr>
<tr>
<td>each = some</td>
<td>656</td>
<td>647 (99%)</td>
</tr>
<tr>
<td>all = many</td>
<td>873</td>
<td>772 (88%)</td>
</tr>
<tr>
<td>much = some</td>
<td>248</td>
<td>217 (88%)</td>
</tr>
<tr>
<td>every = many</td>
<td>460</td>
<td>400 (87%)</td>
</tr>
<tr>
<td>many = some</td>
<td>951</td>
<td>822 (86%)</td>
</tr>
<tr>
<td>all = most</td>
<td>465</td>
<td>393 (85%)</td>
</tr>
<tr>
<td>several = some</td>
<td>580</td>
<td>439 (76%)</td>
</tr>
<tr>
<td>both = some</td>
<td>573</td>
<td>322 (56%)</td>
</tr>
<tr>
<td>many = several</td>
<td>594</td>
<td>113 (19%)</td>
</tr>
<tr>
<td>most = many</td>
<td>463</td>
<td>84 (18%)</td>
</tr>
<tr>
<td>both = either</td>
<td>63</td>
<td>1 (2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantifier pair</th>
<th>Instances</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>some ≠ every</td>
<td>484</td>
<td>481 (99%)</td>
</tr>
<tr>
<td>several ≠ all</td>
<td>557</td>
<td>553 (99%)</td>
</tr>
<tr>
<td>several ≠ every</td>
<td>378</td>
<td>375 (99%)</td>
</tr>
<tr>
<td>some ≠ all</td>
<td>1054</td>
<td>1043 (99%)</td>
</tr>
<tr>
<td>many ≠ every</td>
<td>460</td>
<td>452 (98%)</td>
</tr>
<tr>
<td>some ≠ each</td>
<td>656</td>
<td>640 (98%)</td>
</tr>
<tr>
<td>few ≠ all</td>
<td>157</td>
<td>153 (97%)</td>
</tr>
<tr>
<td>many ≠ all</td>
<td>873</td>
<td>843 (97%)</td>
</tr>
<tr>
<td>both ≠ most</td>
<td>369</td>
<td>347 (94%)</td>
</tr>
<tr>
<td>several ≠ few</td>
<td>143</td>
<td>134 (94%)</td>
</tr>
<tr>
<td>both ≠ many</td>
<td>541</td>
<td>397 (73%)</td>
</tr>
<tr>
<td>many ≠ most</td>
<td>463</td>
<td>300 (65%)</td>
</tr>
<tr>
<td>either ≠ both</td>
<td>63</td>
<td>39 (62%)</td>
</tr>
<tr>
<td>many ≠ no</td>
<td>714</td>
<td>369 (52%)</td>
</tr>
<tr>
<td>some ≠ many</td>
<td>951</td>
<td>468 (49%)</td>
</tr>
<tr>
<td>few ≠ many</td>
<td>161</td>
<td>33 (20%)</td>
</tr>
<tr>
<td>both ≠ several</td>
<td>431</td>
<td>63 (15%)</td>
</tr>
</tbody>
</table>