# Entailment above the word level in distributional semantics


Marco Baroni University of Trento
Raffaella Bernardi University of Trento
Ngoc-Quynh Do EM LCT, Free University of Bozen-Bolzano
Chung-chieh Shan Cornell University, University of Tsukuba



EACL
25 April 2012

# 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.)

# 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.)

| **AN $\models$ N** | $\xrightarrow{\text{train} \qquad \text{test}}$ | **N $\models$ N** |
|---|---|---|
| big cat — cat | | dog — animal |

# 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.)

$$\underset{\text{big cat}}{\mathbf{AN}} \models \underset{\text{cat}}{\mathbf{N}} \quad \longrightarrow \quad \underset{\text{dog}}{\mathbf{N}} \models \underset{\text{animal}}{\mathbf{N}}$$

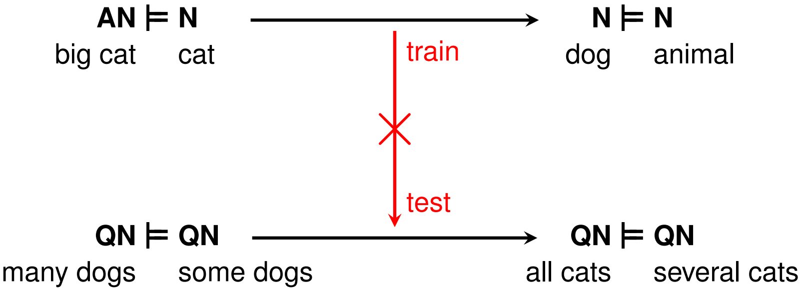
$$\underset{\text{many dogs}}{\mathbf{QN}} \models \underset{\text{some dogs}}{\mathbf{QN}} \quad \xrightarrow{\text{train} \qquad\qquad \text{test}} \quad \underset{\text{all cats}}{\mathbf{QN}} \models \underset{\text{several cats}}{\mathbf{QN}}$$

# 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.)



# 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. | $\vDash$ | Every professor cried. |
| A person cried. | $\nvDash$ | A professor cried. |

## Formal semantics

$$\forall x.\, Px \to Cx$$

$$\lambda g.\, \forall x.\, Px \to gx \qquad C$$

$$\lambda f.\, \lambda g.\, \forall x.\, fx \to gx \qquad P$$

# 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

$$\begin{array}{lcl} \text{ambulance} & \sim & \text{battleship} \\ \text{ambulance} & \nsim & \text{bookstore} \end{array}$$

## Distributional semantics

$$\begin{array}{l} \\ \text{ambulance} \\ \text{battleship} \\ \text{bookstore} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \text{accept} & \dots \end{array} \\ \begin{pmatrix} 27 & 10 & 50 & 17 & 130 & \dots \\ 35 & 0 & 32 & 1 & 25 & \dots \\ 5 & 0 & 6 & 33 & 13 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$


衆瞽
探象之圖

# Distributional semantics for entailment among words

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

"pointwise mutual information"

# Distributional semantics for entailment among words

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

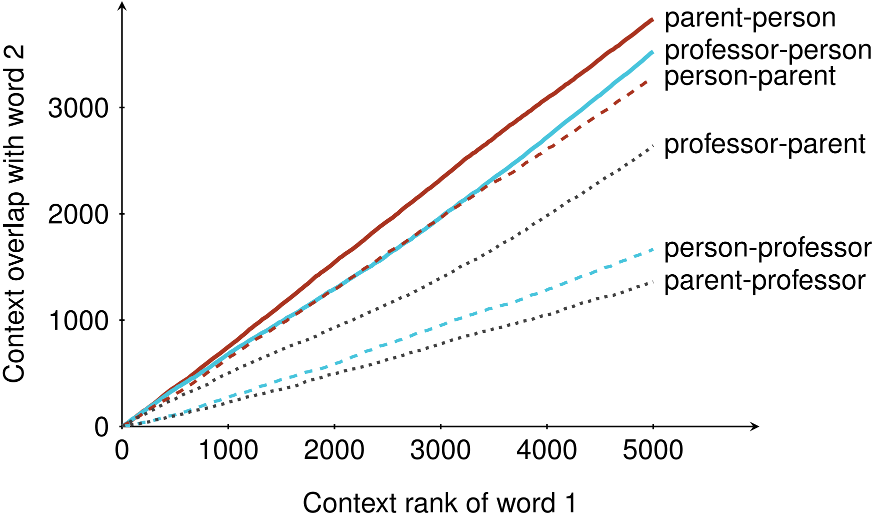
"pointwise mutual information"

**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$_n$ symptons$_n$ adoptive$_j$ stepparent$_n$ nonresident$_j$ home-school$_n$ scabrid$_n$ petiolule$_n$ ...

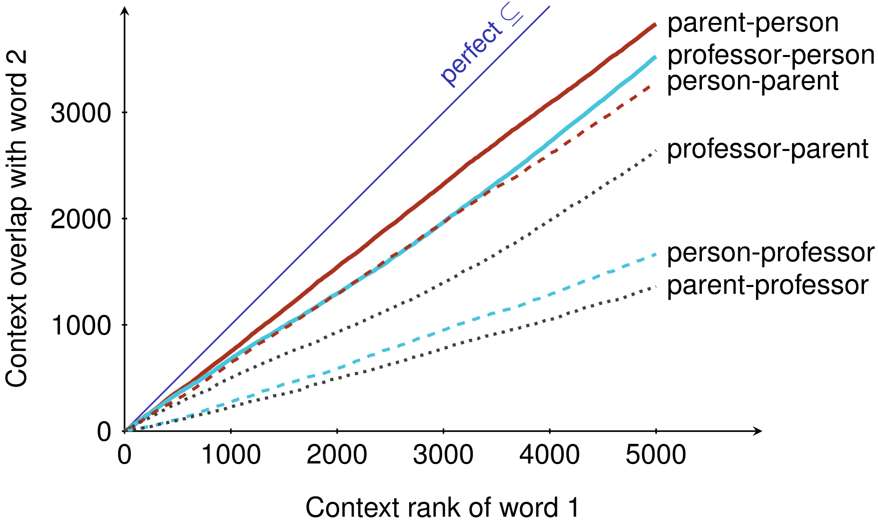
**person** anglia$_n$ first-mentioned$_j$ unascertained$_j$ enure$_v$ deposit-taking$_j$ bonis$_n$ iconclass$_j$ cotswolds$_n$ aforesaid$_n$ haver$_v$ foresaid$_j$ gha$_n$ sub-paragraphs$_n$ enacted$_j$ geest$_j$ non-medicinal$_j$ sub-paragraph$_n$ intimation$_n$ arrestment$_n$ incumbrance$_n$ ...

**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$_n$ ...

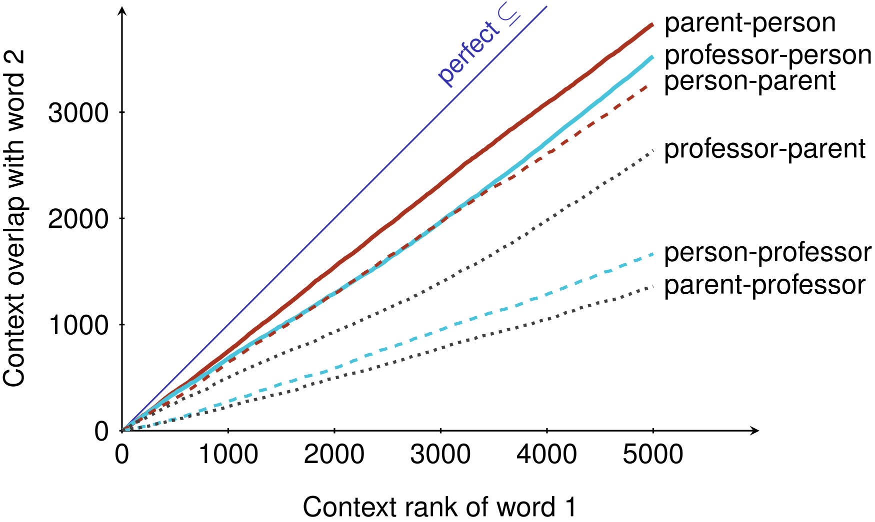
# Distributional semantics for entailment among words



# Distributional semantics for entailment among words



# Distributional semantics for entailment among words




Better: *skew divergence* (Lee), *balAPinc* (Kotlerman et al.), . . .

# 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**

| | | Syntactic category |
|---|---|---|
| **N** | cat | N |
| **AN** | white cat | N |
| **QN** | every cat | QP |

# Above the word level

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

| | | Syntactic category | Semantic type |
|---|---|---|---|
| **N** | cat | N | $e \to t$ |
| **AN** | white cat | N | $e \to t$ |
| **QN** | every cat | QP | $(e \to t) \to t$ |

# Above the word level

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

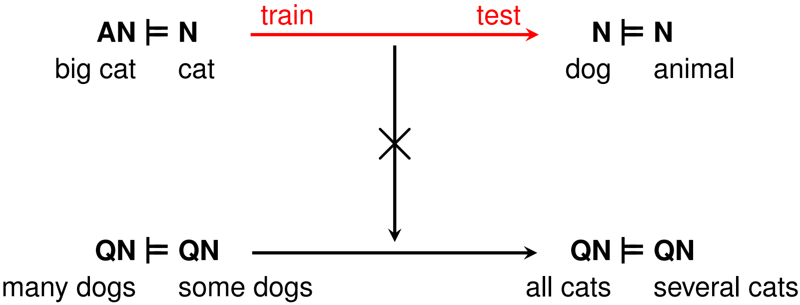
| | | Syntactic category | Semantic type |
|---|---|---|---|
| **N** | cat | N | $e \to t$ |
| **AN** | white cat | N | $e \to t$ |
| **AAN** | big white cat | N | $e \to t$ |
| **QN** | every cat | QP | $(e \to t) \to t$ |
| **QAN** | every big cat | QP | $(e \to t) \to t$ |
| * **AQN** | big every cat | | |
| * **QQN** | some every cat | | |

# Our questions

Entailment among composite phrases rather than nouns?

Entailment among logical words rather than content words?

Different entailment relations at different semantic types?

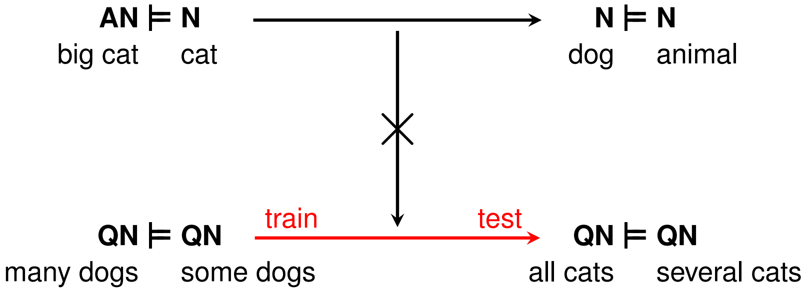


# Our questions

Entailment among composite phrases rather than nouns?

Entailment among logical words rather than content words?

Different entailment relations at different semantic types?



# Our questions

Entailment among composite phrases rather than nouns?

Entailment among logical words rather than content words?

Different entailment relations at different semantic types?

$\mathbf{AN} \models_{\mathbf{N}} \mathbf{N}$ $\longrightarrow$ $\mathbf{N} \models_{\mathbf{N}} \mathbf{N}$

big cat $\models_{\mathbf{N}}$ cat — dog $\models_{\mathbf{N}}$ animal

train

test

$\mathbf{QN} \models_{\mathbf{QN}} \mathbf{QN}$ $\longrightarrow$ $\mathbf{QN} \models_{\mathbf{QN}} \mathbf{QN}$

many dogs $\models_{\mathbf{QN}}$ some dogs — all cats $\models_{\mathbf{QN}}$ several cats

# Our semantic space

BNC, WackyPedia, ukWaC

$\downarrow$ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

$\downarrow$ words and phrases in the same sentence

most frequent
A, N, V (27K)

$$\begin{array}{r}\text{AN}\\ \text{QN}\\ \text{A}\\ \text{Q}\\ \text{N}\\ \text{(48K)}\end{array}\left(\begin{array}{c}\\ \\ \quad \#(c, w) \quad \\ \\ \\ \end{array}\right)$$

# Our semantic space

BNC, WackyPedia, ukWaC

$\downarrow$ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

$\downarrow$ words and phrases in the same sentence

$$\begin{array}{rc} & \begin{array}{c}\text{most frequent}\\ \text{A, N, V (27K)}\end{array} \\ \begin{array}{r}\text{AN}\\ \text{QN}\\ \text{A}\\ \text{Q}\\ \text{N}\\ \text{(48K)}\end{array} & \left(\begin{array}{c} \\ \#(c, w) \\ \\ \end{array}\right) \end{array} \xrightarrow{\text{PMI}} \left(\log \frac{\Pr(c|w)}{\Pr(c)}\right) \xrightarrow{\text{SVD}} \begin{array}{c}\text{(300)}\\ \left(\begin{array}{c} \\ U\tilde{\Sigma} \\ \\ \end{array}\right)\end{array}$$

# Our semantic space

BNC, WackyPedia, ukWaC

$\downarrow$ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

$\downarrow$ words and phrases in the same sentence

$$
\begin{array}{r}
\\ \text{AN} \\ \text{QN} \\ \text{A} \\ \text{Q} \\ \text{N} \\ (48\text{K})
\end{array}
\overset{\substack{\text{most frequent} \\ \text{A, N, V (27K)}}}{\left(\quad \#(c, w) \quad\right)}
\xrightarrow{\text{PMI}}
\left( \log \frac{\Pr(c|w)}{\Pr(c)} \right)
\xrightarrow{\text{SVD}}
\overset{(300)}{\left( U\tilde{\Sigma} \right)}
$$

$\downarrow$ frequency baseline

$\downarrow$ cosine baseline

$\downarrow$ balAPinc

$\downarrow$ SVM

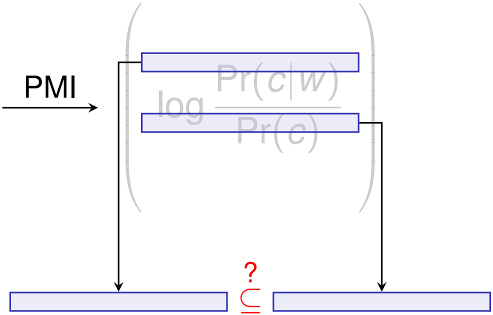
# Our entailment classifiers

$$\xrightarrow{\text{PMI}} \left( \log \frac{\Pr(c|w)}{\Pr(c)} \right)$$

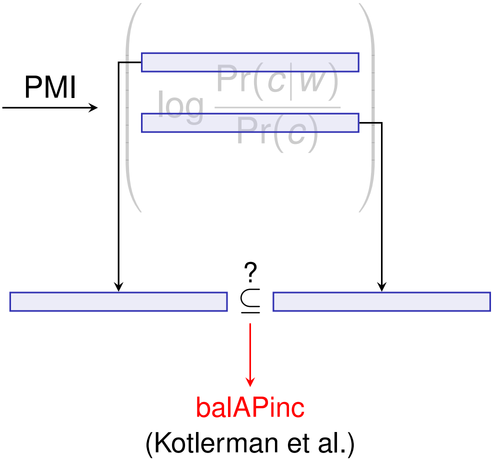
# Our entailment classifiers

$$\xrightarrow{\text{PMI}} \left( \log \frac{\Pr(c|w)}{\Pr(c)} \right)$$

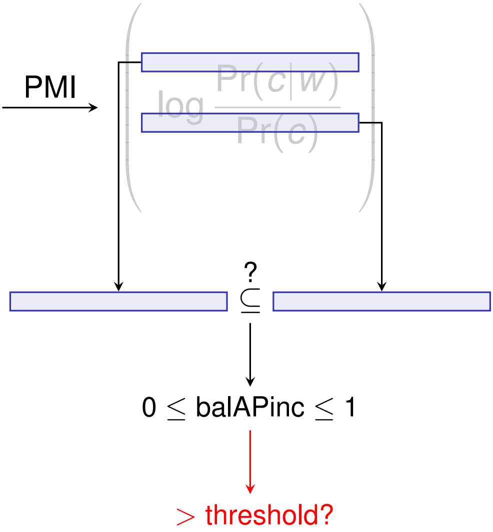
# Our entailment classifiers



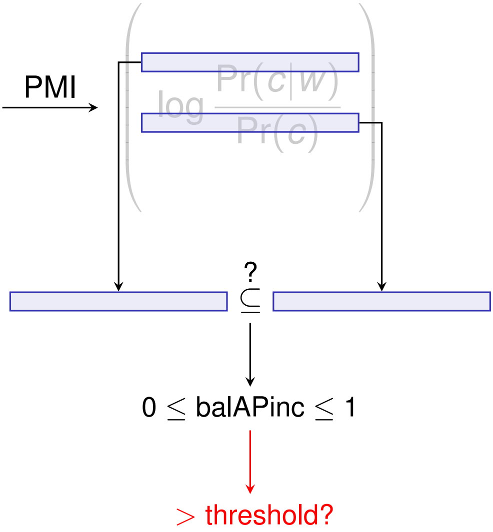
# Our entailment classifiers



# Our entailment classifiers

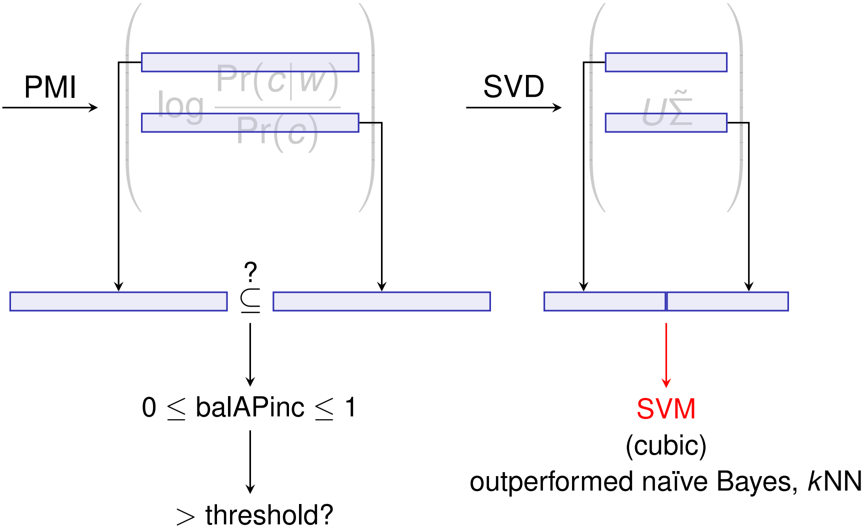


# Our entailment classifiers




| Train | Test |
|---|---|
| AN $\vDash$ N | N $\vDash$ N |
| QN $\vDash$ QN | QN $\vDash$ QN |
| AN $\vDash$ N | QN $\vDash$ QN |

# Our entailment classifiers



# Our data sets

WordNet

$$\downarrow$$

$$\begin{aligned}
\text{pope} &\models \text{spiritual_leader} \\
\text{spiritual_leader} &\models \text{leader} \\
\text{cat} &\models \text{feline} \\
\text{feline} &\models \text{carnivore} \\
&\vdots
\end{aligned}$$

# Our data sets

WordNet

$\downarrow$

pope $\models$ leader
cat $\models$ carnivore
$\vdots$ (1385)

# Our data sets

$$
\begin{tikzcd}
\text{WordNet} \arrow[d] \\
\begin{array}{rcl} \text{pope} & \models & \text{leader} \\ \text{cat} & \models & \text{carnivore} \\ & \vdots & (1385) \end{array} \arrow[d, bend right, "\text{invert}"'] \arrow[d, bend left, "\text{resample}"] \\
\begin{array}{rcl} \text{leader} & \not\models & \text{pope} \\ \text{cat} & \not\models & \text{leader} \\ & \vdots & (1385) \end{array}
\end{tikzcd}
$$

# Our data sets

most frequent

$\downarrow$

big
former
$\vdots$ (300)

WordNet

$\downarrow$

pope $\models$ leader
cat $\models$ carnivore
$\vdots$ (1385)

invert

resample

leader $\not\models$ pope
cat $\not\models$ leader
$\vdots$ (1385)

# Our data sets

most frequent

↓

big

~~former~~

⋮ (256)

WordNet

↓

pope $\models$ leader

cat $\models$ carnivore

⋮ (1385)

invert

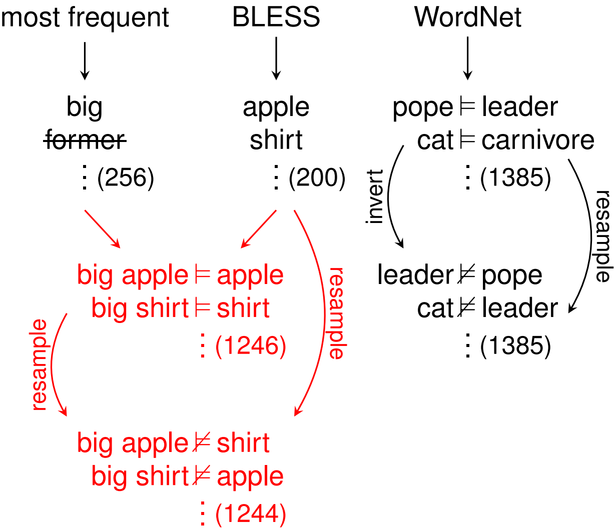
resample

leader $\not\models$ pope

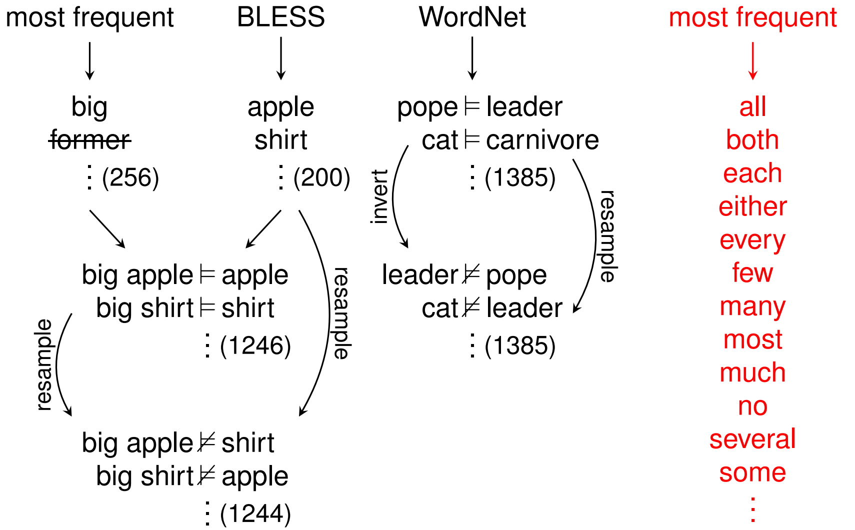
cat $\not\models$ leader

⋮ (1385)

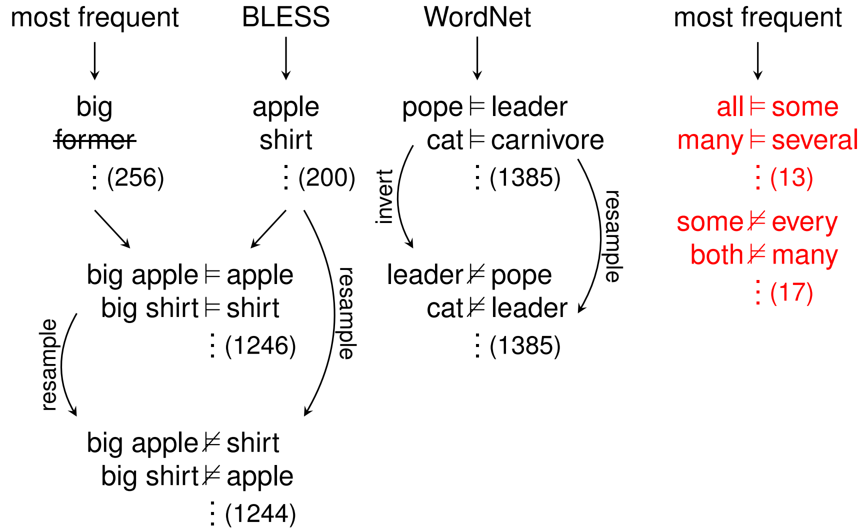
# Our data sets



# Our data sets



# Our data sets



# Our data sets

most frequent

big
~~former~~
⋮ (256)

BLESS

apple
shirt
⋮ (200)

big apple $\vDash$ apple
big shirt $\vDash$ shirt
⋮ (1246)

resample

resample

big apple $\nvDash$ shirt
big shirt $\nvDash$ apple
⋮ (1244)

WordNet

pope $\vDash$ leader
cat $\vDash$ carnivore
⋮ (1385)

invert

resample

leader $\nvDash$ pope
cat $\nvDash$ leader
⋮ (1385)

pope
leader
cat
carnivore
⋮ (6402)

most frequent

all $\vDash$ some
many $\vDash$ several
⋮ (13)

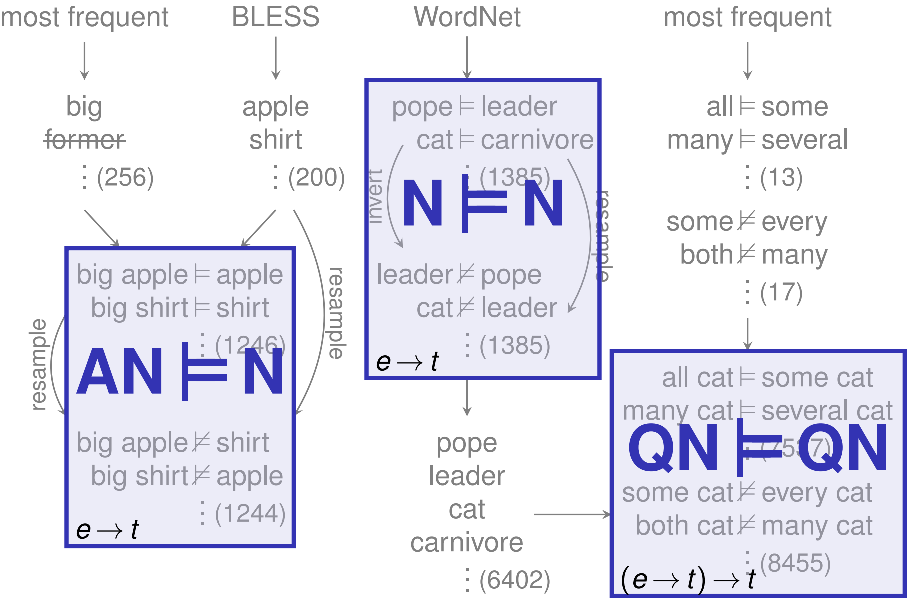
some $\nvDash$ every
both $\nvDash$ many
⋮ (17)

all cat $\vDash$ some cat
many cat $\vDash$ several cat
⋮ (7537)

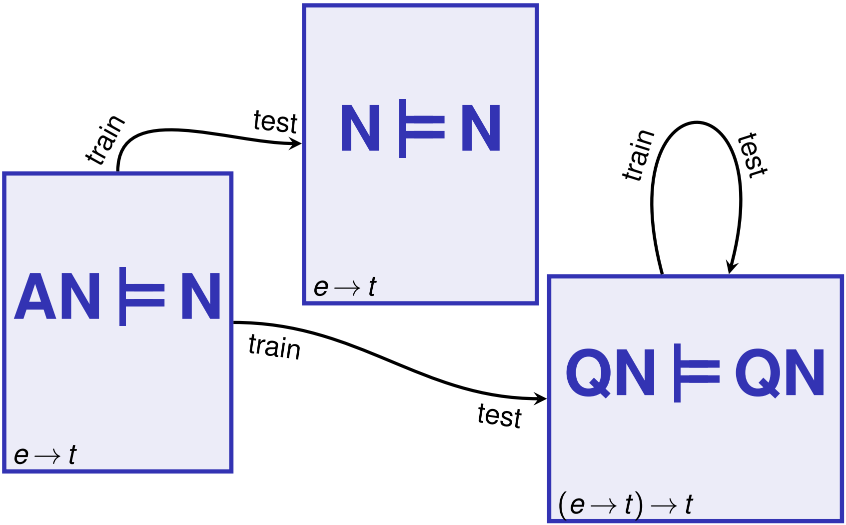
some cat $\nvDash$ every cat
both cat $\nvDash$ many cat
⋮ (8455)

all cat $\nvDash$ every leader
both cat $\nvDash$ many leader

# Our data sets



# Our data sets



# Results at noun type

| | P | R | F | Accuracy | (95% C.I.) |
|---|---|---|---|---|---|
| $\mathrm{SVM}_{\mathrm{upper}}$ | 88.6 | 88.6 | 88.5 | 88.6 | (87.3–89.7) |
| $\mathrm{balAPinc}_{\mathrm{AN} \models \mathrm{N}}$ | 65.2 | 87.5 | 74.7 | 70.4 | (68.7–72.1) |
| $\mathrm{balAPinc}_{\mathrm{upper}}$ | 64.4 | 90.0 | 75.1 | 70.1 | (68.4–71.8) |
| $\mathrm{SVM}_{\mathrm{AN} \models \mathrm{N}}$ | 69.3 | 69.3 | 69.3 | 69.3 | (67.6–71.0) |
| $\cos(\mathrm{N}_1, \mathrm{N}_2)$ | 57.7 | 57.6 | 57.5 | 57.6 | (55.8–59.5) |
| $\mathrm{fq}(\mathrm{N}_1) < \mathrm{fq}(\mathrm{N}_2)$ | 52.1 | 52.1 | 51.8 | 53.3 | (51.4–55.2) |

# Holding out QN data

| $\models$ | all | both | each | either | every | few | many | most | much | no | several | some |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | | | | | | | + | + | | | + | + |
| both | | | | + | | | − | − | | | − | + |
| each | | | | | | | | | | | | + |
| either | | − | | | | | | | | | | |
| every | | | | | | | + | | | | | |
| few | − | | | | | | − | | | | | |
| many | − | | | | − | | | − | | − | + | + |
| most | | | | | | | + | | | | | |
| much | | | | | | | | | | | | + |
| no | | | | | | | | | | | | |
| several | − | | | | − | − | | | | | | + |
| some | − | | − | | − | | − | | | | | |

# Holding out QN data

| $\models$ | all | both | each | either | every | few | many | most | much | no | several | some |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | | | | | | | + | + | | | + | + |
| both | | | | + | | | − | − | | | − | + |
| each | | | | | | | | | | | | + |
| either | | − | | | | | | | | | | |
| every | | | | | | | + | | | | | |
| few | − | | | | | | − | | | | | |
| many | − | | | | − | | | − | | − pair-out | + | + |
| most | | | | | | | + | | | | | |
| much | | | | | | | | | | | | + |
| no | | | | | | | | | | | | |
| several | − | | | | − | − | | | | | | + |
| some | − | | − | | − | | − | | | | | |

# Holding out QN data

| ⊭ | all | both | each | either | every | few | many | most | much | no | several | some |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | | | | | | | + | + | | | + | + |
| both | | | | + | | | − | − | | | − | + |
| each | | | | | | | | | | | | + |
| either | | − | | | | | | | | | | |
| every | | | | | | | + | | | | | |
| few | − | | | | | | − | | | | | |
| many | − | | | | − | | | − | | − | + | + |
| most | | | | | | | + | | | | | |
| much | | | | | | | | | | | | + |
| no | | | | | | | | | | | | |
| several | − | | | | − | − | | | | | | + |
| some | − | | − | | − | | − | | | | | |

pair-out

quantifier-out

# Results at quantifier type

| | P | R | F | Accuracy | (95% C.I.) |
|---|---|---|---|---|---|
| $\mathrm{SVM}_{\text{pair-out}}$ | 76.7 | 77.0 | 76.8 | 78.1 | (77.5–78.8) |
| $\mathrm{SVM}_{\text{quantifier-out}}$ | 70.1 | 65.3 | 68.0 | 71.0 | (70.3–71.7) |
| $\mathrm{SVM}^{Q}_{\text{pair-out}}$ | 67.9 | 69.8 | 68.9 | 70.2 | (69.5–70.9) |
| $\mathrm{SVM}^{Q}_{\text{quantifier-out}}$ | 53.3 | 52.9 | 53.1 | 56.0 | (55.2–56.8) |
| $\cos(QN_1, QN_2)$ | 52.9 | 52.3 | 52.3 | 53.1 | (52.3–53.9) |
| $\mathrm{balAPinc}_{AN \models N}$ | 46.7 | 5.6 | 10.0 | 52.5 | (51.7–53.3) |
| $\mathrm{SVM}_{AN \models N}$ | 2.8 | 42.9 | 5.2 | 52.4 | (51.7–53.2) |
| $fq(QN_1) < fq(QN_2)$ | 51.0 | 47.4 | 49.1 | 50.2 | (49.4–51.0) |
| $\mathrm{balAPinc}_{\text{upper}}$ | 47.1 | 100 | 64.1 | 47.2 | (46.4–47.9) |

# Holding out each quantifier

| Quantifier | Instances | | Correct | | |
|---|---|---|---|---|---|
| | $\models$ | $\not\models$ | $\models$ | $\not\models$ | |
| each | 656 | 656 | 649 | 637 | (98%) |
| every | 460 | 1322 | 402 | 1293 | (95%) |
| much | 248 | 0 | 216 | 0 | (87%) |
| all | 2949 | 2641 | 2011 | 2494 | (81%) |
| several | 1731 | 1509 | 1302 | 1267 | (79%) |
| many | 3341 | 4163 | 2349 | 3443 | (77%) |
| few | 0 | 461 | 0 | 311 | (67%) |
| most | 928 | 832 | 549 | 511 | (60%) |
| some | 4062 | 3145 | 1780 | 2190 | (55%) |
| no | 0 | 714 | 0 | 380 | (53%) |
| both | 636 | 1404 | 589 | 303 | (44%) |
| either | 63 | 63 | 2 | 41 | (34%) |
| *Total* | *15074* | *16910* | *9849* | *12870* | *(71%)* |

# Our questions answered

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

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

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

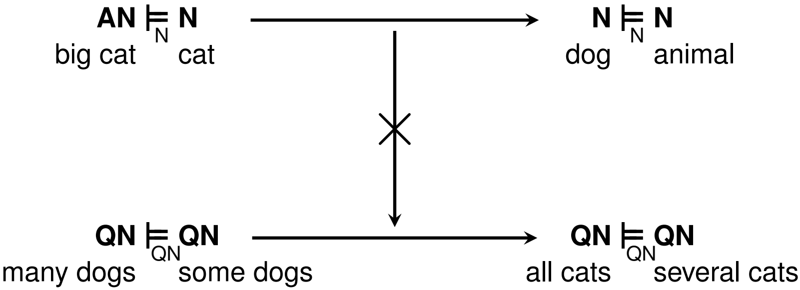
$$\begin{tikzcd}[column sep=huge, row sep=huge]
\begin{array}{c}\mathbf{AN} \models_{\mathrm{N}} \mathbf{N} \\ \text{big cat} \quad \text{cat}\end{array} \arrow[r] \arrow[d, phantom, ""] & \begin{array}{c}\mathbf{N} \models_{\mathrm{N}} \mathbf{N} \\ \text{dog} \quad \text{animal}\end{array} \\
\begin{array}{c}\mathbf{QN} \models_{\mathrm{QN}} \mathbf{QN} \\ \text{many dogs} \quad \text{some dogs}\end{array} \arrow[r] \arrow[u, phantom, shift left=20ex, "\not\Downarrow"] & \begin{array}{c}\mathbf{QN} \models_{\mathrm{QN}} \mathbf{QN} \\ \text{all cats} \quad \text{several cats}\end{array}
\end{tikzcd}$$

# Our questions answered

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

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

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



# Our questions answered

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

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

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

Ongoing work:

- How does the SVM work?
- Missing experiments?
- How to compose semantic vectors?

# Holding out each quantifier pair

| Quantifier pair | Instances | Correct |
|---|---|---|
| all $\models$ some | 1054 | 1044 (99%) |
| all $\models$ several | 557 | 550 (99%) |
| each $\models$ some | 656 | 647 (99%) |
| all $\models$ many | 873 | 772 (88%) |
| much $\models$ some | 248 | 217 (88%) |
| every $\models$ many | 460 | 400 (87%) |
| many $\models$ some | 951 | 822 (86%) |
| all $\models$ most | 465 | 393 (85%) |
| several $\models$ some | 580 | 439 (76%) |
| both $\models$ some | 573 | 322 (56%) |
| many $\models$ several | 594 | 113 (19%) |
| most $\models$ many | 463 | 84 (18%) |
| both $\models$ either | 63 | 1 (2%) |

| Quantifier pair | Instances | Correct |
|---|---|---|
| some $\not\models$ every | 484 | 481 (99%) |
| several $\not\models$ all | 557 | 553 (99%) |
| several $\not\models$ every | 378 | 375 (99%) |
| some $\not\models$ all | 1054 | 1043 (99%) |
| many $\not\models$ every | 460 | 452 (98%) |
| some $\not\models$ each | 656 | 640 (98%) |
| few $\not\models$ all | 157 | 153 (97%) |
| many $\not\models$ all | 873 | 843 (97%) |
| both $\not\models$ most | 369 | 347 (94%) |
| several $\not\models$ few | 143 | 134 (94%) |
| both $\not\models$ many | 541 | 397 (73%) |
| many $\not\models$ most | 463 | 300 (65%) |
| either $\not\models$ both | 63 | 39 (62%) |
| many $\not\models$ no | 714 | 369 (52%) |
| some $\not\models$ many | 951 | 468 (49%) |
| few $\not\models$ many | 161 | 33 (20%) |
| both $\not\models$ several | 431 | 63 (15%) |