

Natural logic and natural language inference



Bill MacCartney
BlackRock / Stanford
10 April 2015

Motivating examples

- P. A Revenue Cutter, the ship was named for Harriet Lane, niece of President James Buchanan, who served as Buchanan's White House hostess.
- H. Harriet Lane worked at the White House. **yes**
- P. Two Turkish engineers and an Afghan translator kidnapped in July were freed Friday.
- H. translator kidnapped in Iraq **no**
- P. The memorandum noted the United Nations estimated that 2.5 million to 3.5 million people died of AIDS last year.
- H. Over 2 million people died of AIDS last year. **yes**
- P. Mitsubishi Motors Corp.'s new vehicle sales in the US fell 46 percent in June.
- H. Mitsubishi sales rose 46 percent. **no**
- P. The main race track in Qatar is located in Shahaniya, on the Dukhan Road.
- H. Qatar is located in Shahaniya. **no**

Natural language inference (NLI)

- Does premise P justify an inference to hypothesis H ?
 - An informal, intuitive notion of inference: not strict logic
 - Focus on local inference steps, not long chains of deduction
 - Emphasis on variability of linguistic expression
- Robust, accurate natural language inference could enable:
 - Semantic search
 - H: *lobbyists attempting to bribe U.S. legislators*
 - P: *The A.P. named two more senators who received contributions engineered by lobbyist Jack Abramoff in return for political favors*
 - Question answering [Harabagiu & Hickl 06]
 - H: *Who bought JDE?* P: *Thanks to its recent acquisition of JDE, Oracle will ...*
 - Document summarization
- Cf. paraphrase task: do sentences P and Q mean the same?
 - natural language inference: $P \rightarrow Q$ Paraphrase: $P \leftrightarrow Q$

NLI and NLU

- The ability to draw simple inferences is a key test of understanding
 - P. *The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.*
 - H. *Jill Carroll was abducted in Iraq.*
- If you can't recognize that P implies H, then you haven't really understood P (or H)
- Thus, a capacity for natural language inference is a necessary (though probably not sufficient) condition for real NLU

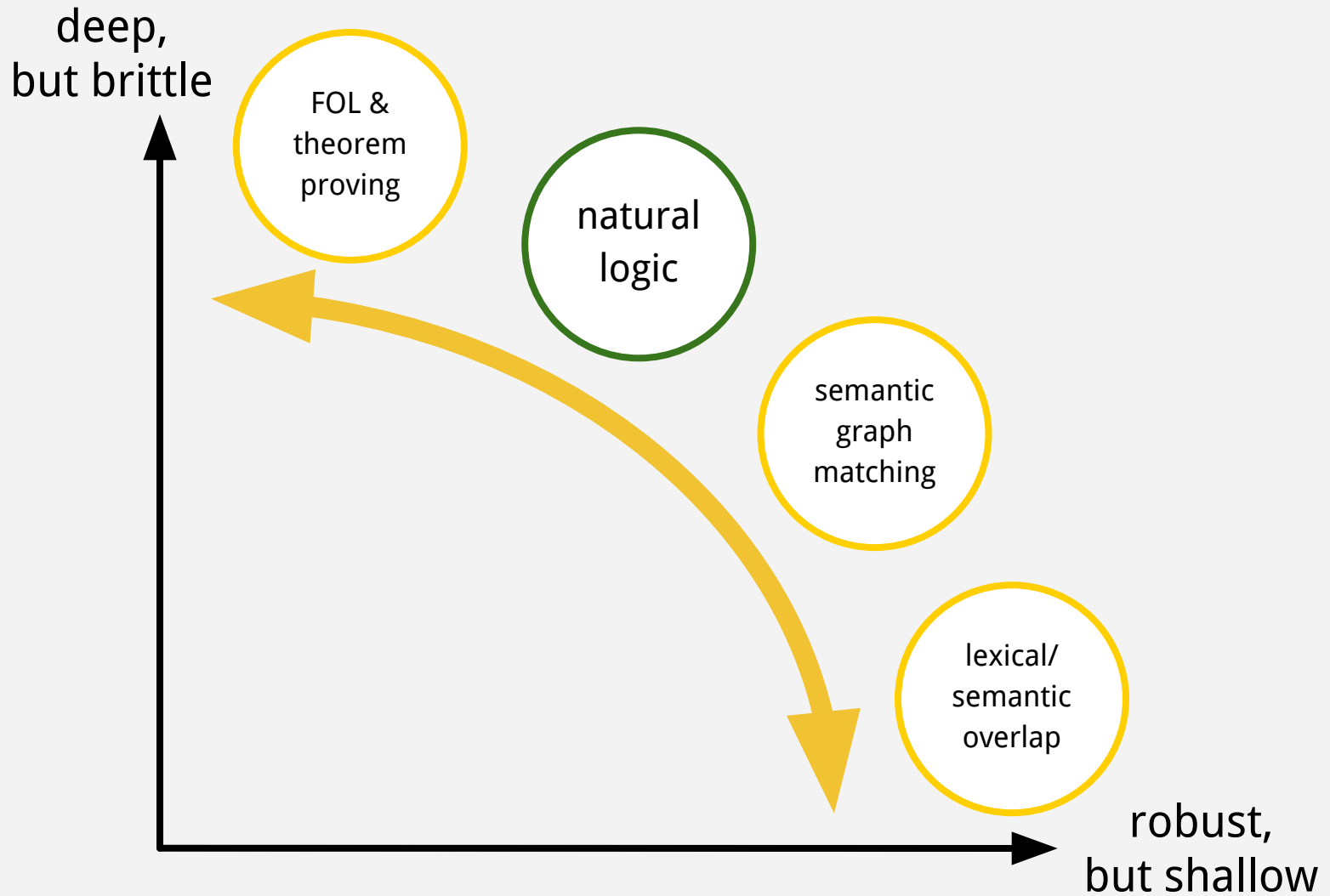
The RTE challenges

- RTE = Recognizing Textual Entailment
- Eight annual competitions: RTE-1 (2005) to RTE-8 (2013)
- Typical data sets: 800 training pairs, 800 test pairs
- Earlier competitions were binary decision tasks
 - Entailment vs. no entailment
- Three-way decision task introduced with RTE-4
 - Entailment, contradiction, unknown
- Lots of resources available:
http://aclweb.org/aclwiki/index.php?title=Textual_Entailment

The SICK dataset

- SICK = Sentences Involving Compositional Knowledge
- The basis of a shared task in SEMEVAL 2014
- 10,000 sentence pairs, derived from image and video captions
- Annotated with two labels via crowdsourcing
 - Sentence relatedness: a five-point scale
 - Entailment relation: entailment, contradiction, unknown
- See <http://clic.cimec.unitn.it/composes/sick.html>

Approaches to NLI



Outline

- The natural language inference task
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
 - An algebra of semantic relations
 - An account of compositional entailment
 - A weak proof procedure
- NatLog: implementation & evaluation
- More recent work by others

What is natural logic?

- (natural logic \neq natural deduction)
- Lakoff (1970) defines **natural logic** as a goal (not a system)
 - to characterize valid patterns of reasoning via surface forms (syntactic forms as close as possible to natural language)
 - without translation to formal notation: $\rightarrow \neg \wedge \vee \forall \exists$
- A long history
 - traditional logic: Aristotle's syllogisms, scholastics, Leibniz, ...
 - van Benthem & Sánchez Valencia (1986-91): **monotonicity calculus**
- Precise, yet sidesteps difficulties of translating to FOL:
 - idioms, intensionality and propositional attitudes, modalities, indexicals, reciprocals, scope ambiguities, quantifiers such as *most*, reciprocals, anaphoric adjectives, temporal and causal relations, aspect, unselective quantifiers, adverbs of quantification, donkey sentences, generic determiners, ...

The subsumption principle

- Deleting modifiers & other content (usually) preserves truth
- Inserting new content (usually) does not
- Many approximate approaches to RTE exploit this heuristic
 - Try to match each word or phrase in H to something in P
 - Punish examples which introduce new content in H

P. *The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.*

H. *Jill Carroll was abducted in Iraq.* yes

P. *Two Turkish engineers and an Afghan translator kidnapped in July were freed Friday.*

H. *A translator was kidnapped in Iraq.* no

Upward monotonicity

- Actually, there's a more general principle at work
- Edits which broaden or weaken usually preserve truth

My cat ate a rat \Rightarrow *My cat ate a **rodent***

My cat ate a rat \Rightarrow *My cat **consumed** a rat*

My cat ate a rat this morning \Rightarrow *My cat ate a rat **today***

*My cat ate a **fat** rat* \Rightarrow *My cat ate a rat*

- Edits which narrow or strengthen usually do not

My cat ate a rat \nRightarrow *My cat ate a **Norway** rat*

My cat ate a rat \nRightarrow *My cat ate a rat **with cute little whiskers***

My cat ate a rat last week \nRightarrow *My cat ate a rat last **Tuesday***

Semantic containment

- There are many different ways to broaden meaning!
- Deleting modifiers, qualifiers, adjuncts, appositives, etc.:
tall girl standing by the pool \sqsubset *tall girl* \sqsubset *girl*
- Generalizing instances or classes into superclasses:
Einstein \sqsubset *a physicist* \sqsubset *a scientist*
- Spatial & temporal broadening:
in Palo Alto \sqsubset *in California, this month* \sqsubset *this year*
- Relaxing modals: *must* \sqsubset *could, definitely* \sqsubset *probably* \sqsubset *maybe*
- Relaxing quantifiers: *six* \sqsubset *several* \sqsubset *some*
- Dropping conjuncts, adding disjuncts:
danced and sang \sqsubset *sang* \sqsubset *hummed or sang*

Downward monotonicity

- Certain context elements can *reverse* this heuristic!
- Most obviously, negation

*My cat did **not** eat a rat* \Leftarrow *My cat did not eat a rodent*

- But also many other negative or restrictive expressions!

***No** cats ate rats* \Leftarrow *No cats ate rodents*

***Every** rat fears my cat* \Leftarrow *Every rodent fears my cat*

*My cat ate **at most three** rats* \Leftarrow *My cat ate at most three rodents*

***If** my cat eats a rat, he'll puke* \Leftarrow *If my cat eats a rodent, he'll puke*

*My cat **avoids** eating rats* \Leftarrow *My cat avoids eating rodents*

*My cat **denies** eating a rat* \Leftarrow *My cat denies eating a rodent*

*My cat **rarely** eats rats* \Leftarrow *My cat rarely eats rodents*

Non-monotonicity

- Some context elements block inference in both directions!
- E.g., certain quantifiers, superlatives

Most rats like cheese # Most rodents like cheese

My cat ate exactly three rats # My cat ate exactly three rodents

I climbed the tallest building in Asia # I climbed the tallest building

He is our first black President # He is our first president

Monotonicity calculus (Sánchez Valencia 1991)

- Entailment as semantic containment:
rat \sqsubseteq *rodent*, *eat* \sqsubseteq *consume*, *this morning* \sqsubseteq *today*, *most* \sqsubseteq *some*

- Monotonicity classes for semantic functions
 - Upward monotone: *some rats dream* \sqsubseteq *some rodents dream*
 - Downward monotone: *no rats dream* \supseteq *no rodents dream*
 - Non-monotone: *most rats dream* $\#$ *most rodents dream*

- Handles even nested inversions of monotonicity
Every state forbids shooting game without a hunting license
 + - + - - - + + +

- But lacks any representation of exclusion (negation, antonymy, ...)
Gustav is a dog \sqsubseteq *Gustav is not a Siamese cat*

Semantic exclusion

- Monotonicity calculus deals only with semantic *containment*
- It has nothing to say about semantic *exclusion*
- E.g., negation (exhaustive exclusion)

slept ^ *didn't sleep*

able ^ *unable*

living ^ *nonliving*

sometimes ^ *never*

- E.g., alternation (non-exhaustive exclusion)

cat | *dog*

male | *female*

teacup | *toothbrush*

red | *blue*

hot | *cold*

French | *German*

all | *none*

here | *there*

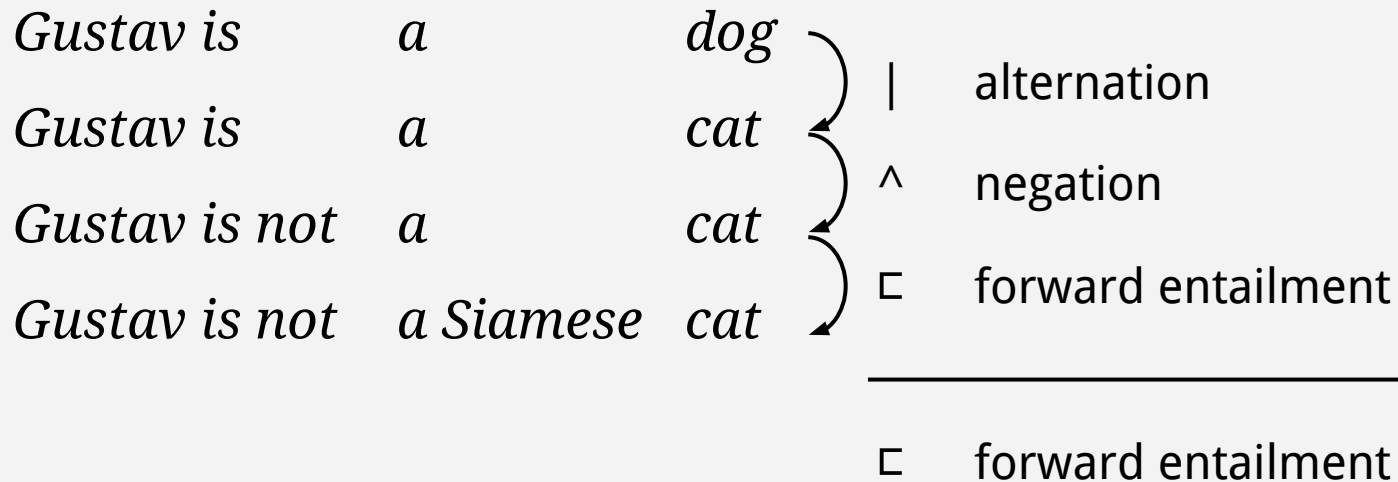
today | *tomorrow*

Outline

- The natural language inference task
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
 - An algebra of semantic relations
 - An account of compositional entailment
 - A weak proof procedure
- NatLog: implementation & evaluation
- More recent work by others

My research agenda, 2007-09

- Build on the **monotonicity calculus** of Sánchez Valencia
- Extend it from semantic containment to **semantic exclusion**
- **Join chains** of semantic containment and exclusion relations
- Apply the system to the task of **natural language inference**



Motivation recap

- To get precise reasoning without full semantic interpretation
 - P. *Every firm surveyed saw costs grow more than expected, even after adjusting for inflation.*
 - H. *Every **big** company in the poll reported cost increases.* **yes**
- Approximate methods fail due to lack of precision
 - Subsumption principle fails — *every* is downward monotone
- Logical methods founder on representational difficulties
 - Full semantic interpretation is difficult, unreliable, expensive
 - How to translate *more than expected* (etc.) to first-order logic?
- Natural logic lets us reason without full interpretation
 - Often, we can drop whole clauses without analyzing them

Semantic relations in past work

<u><i>X is a couch</i></u>	<u><i>X is a crow</i></u>	<u><i>X is a fish</i></u>	<u><i>X is a hippo</i></u>	<u><i>X is a man</i></u>
<i>X is a sofa</i>	<i>X is a bird</i>	<i>X is a carp</i>	<i>X is hungry</i>	<i>X is a woman</i>

2-way
RTE1,2,3

Yes
entailment

No
non-entailment

3-way
RTE4, FraCaS,
PARC, SICK

Yes
entailment

Unknown
non-entailment

No
contradiction

containment
Sánchez-Valencia

$P \equiv Q$
equivalence

$P \sqsubset Q$
forward
entailment

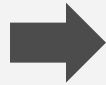
$P \supset Q$
reverse
entailment



$P \# Q$
non-entailment

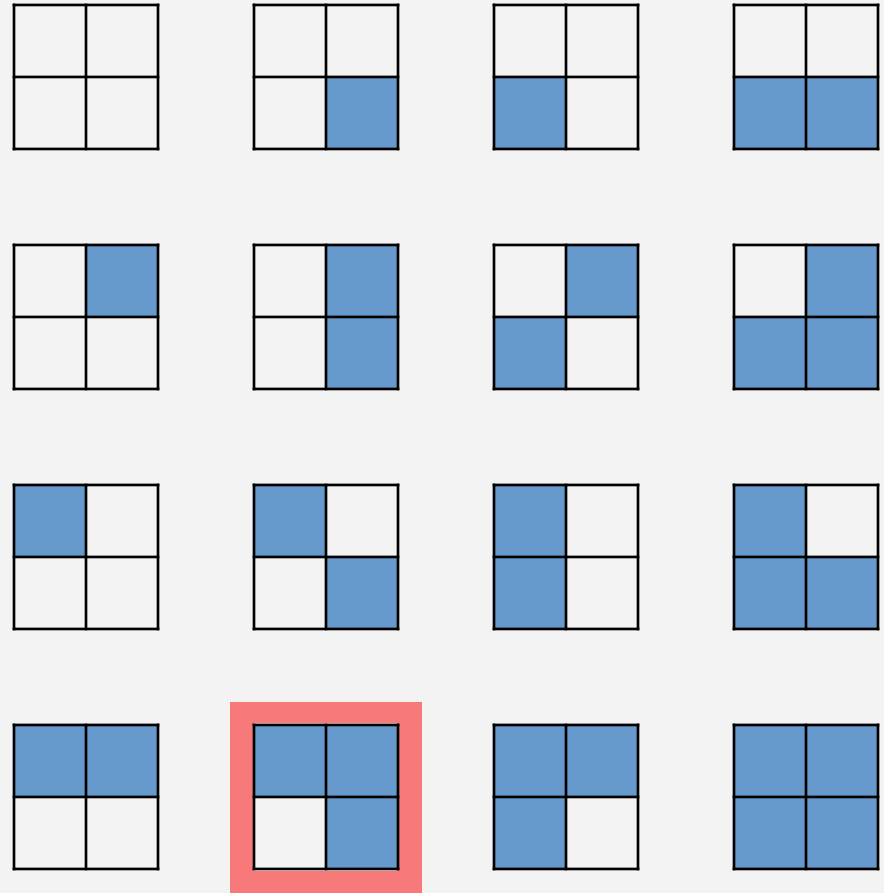
16 elementary set relations

Assign each pair of sets (x, y) to one of 16 relations, depending on the emptiness or non-emptiness of each of the four partitions

	$\neg y$	y
$\neg x$?	?
x	?	?



 empty
 non-empty



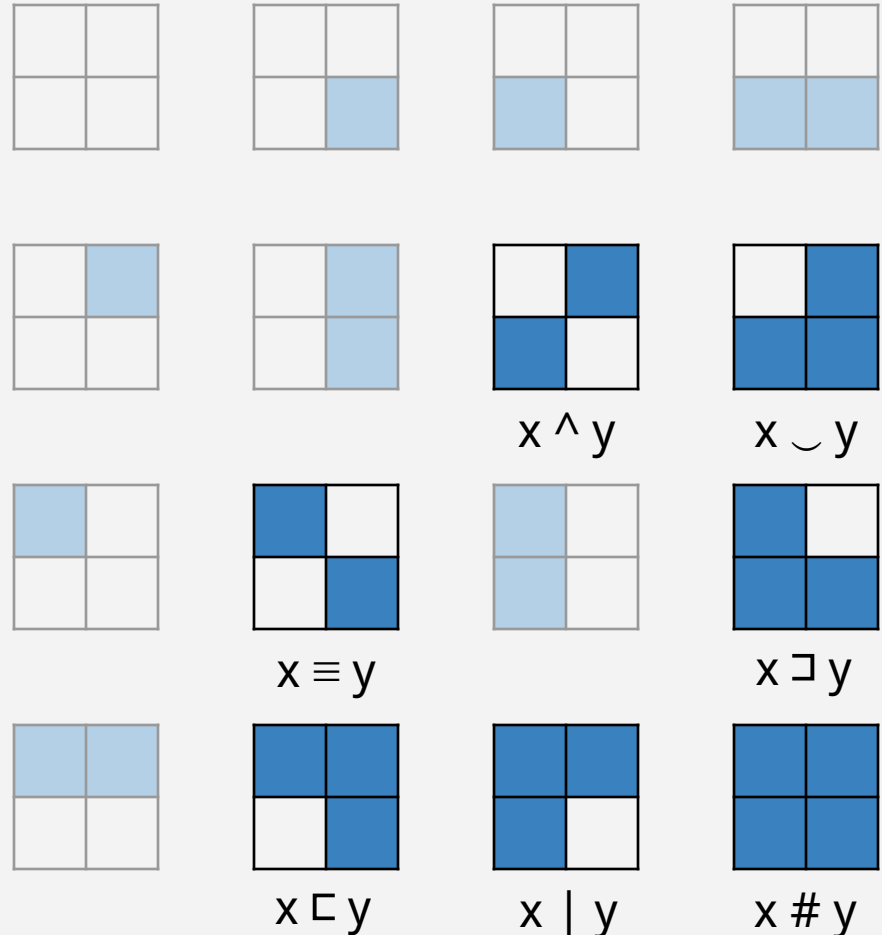
$x \subset y$

16 elementary set relations








But 9 of 16 are degenerate:
either x or y is either empty
or universal.

I.e., they correspond to
semantically vacuous
expressions, which are rare
outside logic textbooks.

We therefore focus on the
remaining seven relations.



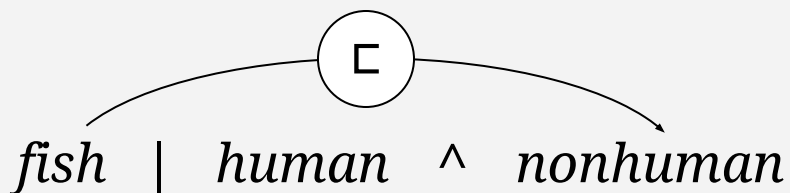
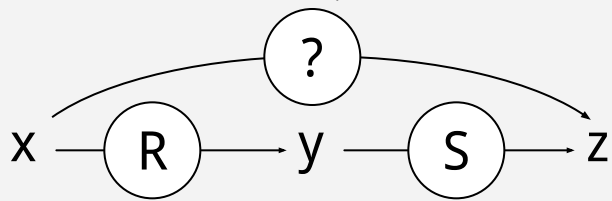
7 basic semantic relations

	$x \equiv y$	equivalence	<i>couch</i> \equiv <i>sofa</i>
	$x \sqsubset y$	forward entailment (strict)	<i>crow</i> \sqsubset <i>bird</i>
	$x \supset y$	reverse entailment (strict)	<i>European</i> \supset <i>French</i>
	$x \wedge y$	negation (exhaustive exclusion)	<i>human</i> \wedge <i>nonhuman</i>
	$x \mid y$	alternation (non-exhaustive exclusion)	<i>cat</i> \mid <i>dog</i>
	$x \smile y$	cover (exhaustive non-exclusion)	<i>animal</i> \smile <i>nonhuman</i>
	$x \# y$	independence	<i>hungry</i> $\#$ <i>hippo</i>

Relations are defined for all semantic types: *tiny* \sqsubset *small*, *hover* \sqsubset *fly*, *kick* \sqsubset *strike*, *this morning* \sqsubset *today*, *in Beijing* \sqsubset *in China*, *everyone* \sqsubset *someone*, *all* \sqsubset *most* \sqsubset *some*

Joining semantic relations

$$R \bowtie S \stackrel{\text{def}}{=} \{ \langle x, z \rangle : \exists y (\langle x, y \rangle \in R \wedge \langle y, z \rangle \in S) \}$$



\equiv	\bowtie	\equiv	\Rightarrow	\equiv
\sqsubset	\bowtie	\sqsubset	\Rightarrow	\sqsubset
\sqsupset	\bowtie	\sqsupset	\Rightarrow	\sqsupset
\wedge	\bowtie	\wedge	\Rightarrow	\equiv
R	\bowtie	\equiv	\Rightarrow	R
\equiv	\bowtie	R	\Rightarrow	R

Some joins yield unions of relations

What is $| \bowtie |$?

$x y$	$y z$	$x ? z$
<i>couch</i> <i>table</i>	<i>table</i> <i>sofa</i>	<i>couch</i> \equiv <i>sofa</i>
<i>pistol</i> <i>knife</i>	<i>knife</i> <i>gun</i>	<i>pistol</i> \sqsubset <i>gun</i>
<i>dog</i> <i>cat</i>	<i>cat</i> <i>terrier</i>	<i>dog</i> \supset <i>terrier</i>
<i>rose</i> <i>orchid</i>	<i>orchid</i> <i>daisy</i>	<i>rose</i> <i>daisy</i>
<i>woman</i> <i>frog</i>	<i>frog</i> <i>Eskimo</i>	<i>woman</i> $\#$ <i>Eskimo</i>

$$| \bowtie | \Rightarrow \cup \{ \equiv, \sqsubset, \supset, |, \# \}$$

The complete join table

\bowtie	\equiv	\sqsubset	\sqsupset	\wedge	\vee	\cup	$\#$
\equiv	\equiv	\sqsubset	\sqsupset	\wedge	\vee	\cup	$\#$
\sqsubset	\sqsubset	\sqsubset	$\equiv \sqsubset \sqsupset \#$	\vee	\sqsupset	$\sqsubset \wedge \cup \#$	$\sqsubset \#$
\sqsupset	\sqsupset	$\equiv \sqsubset \sqsupset \cup \#$	\sqsupset	\cup	$\sqsubset \wedge \cup \#$	\cup	$\sqsupset \cup \#$
\wedge	\wedge	\cup	\vee	\equiv	\sqsupset	\sqsubset	$\#$
\vee	\vee	$\sqsubset \wedge \cup \#$	\vee	\sqsubset	$\equiv \sqsubset \sqsupset \#$	\sqsubset	$\sqsubset \#$
\cup	\cup	\cup	$\sqsubset \wedge \cup \#$	\sqsupset	\sqsupset	$\equiv \sqsubset \sqsupset \cup \#$	$\sqsupset \cup \#$
$\#$	$\#$	$\sqsubset \cup \#$	$\sqsupset \#$	$\#$	$\sqsupset \#$	$\sqsubset \cup \#$	$\equiv \sqsubset \sqsupset \wedge \cup \#$

Of 49 join pairs, 32 yield a single relation; 17 yield unions of relations

Larger unions convey less information — limits power of inference

In practice, any union which contains $\#$ can be approximated by $\#$

Projectivity (= monotonicity++)

- How do the entailments of a compound expression depend on the entailments of its parts?
- How does the semantic relation between $(f x)$ and $(f y)$ depend on the semantic relation between x and y (and the properties of f)?
- Monotonicity gives a partial answer (for $\equiv, \sqsubset, \supset, \#$)
- But what about the other relations (\wedge, \vee, \neg)?
- We'll categorize semantic functions based on how they *project* the basic semantic relations

Example: projectivity of *not*

	projection	example
	$\equiv \rightarrow \equiv$	<i>not happy</i> \equiv <i>not glad</i>
downward monotonicity	$\sqsubset \rightarrow \sqsupset$	<i>didn't kiss</i> \sqsupset <i>didn't touch</i>
	$\sqsupset \rightarrow \sqsubset$	<i>isn't European</i> \sqsubset <i>isn't French</i>
	$\# \rightarrow \#$	<i>isn't swimming</i> $\#$ <i>isn't hungry</i>
	$\wedge \rightarrow \wedge$	<i>not human</i> \wedge <i>not nonhuman</i>
swaps these too	$ \rightarrow \smile$	<i>not French</i> \smile <i>not German</i>
	$\smile \rightarrow $	<i>not more than 4</i> $ $ <i>not less than 6</i>

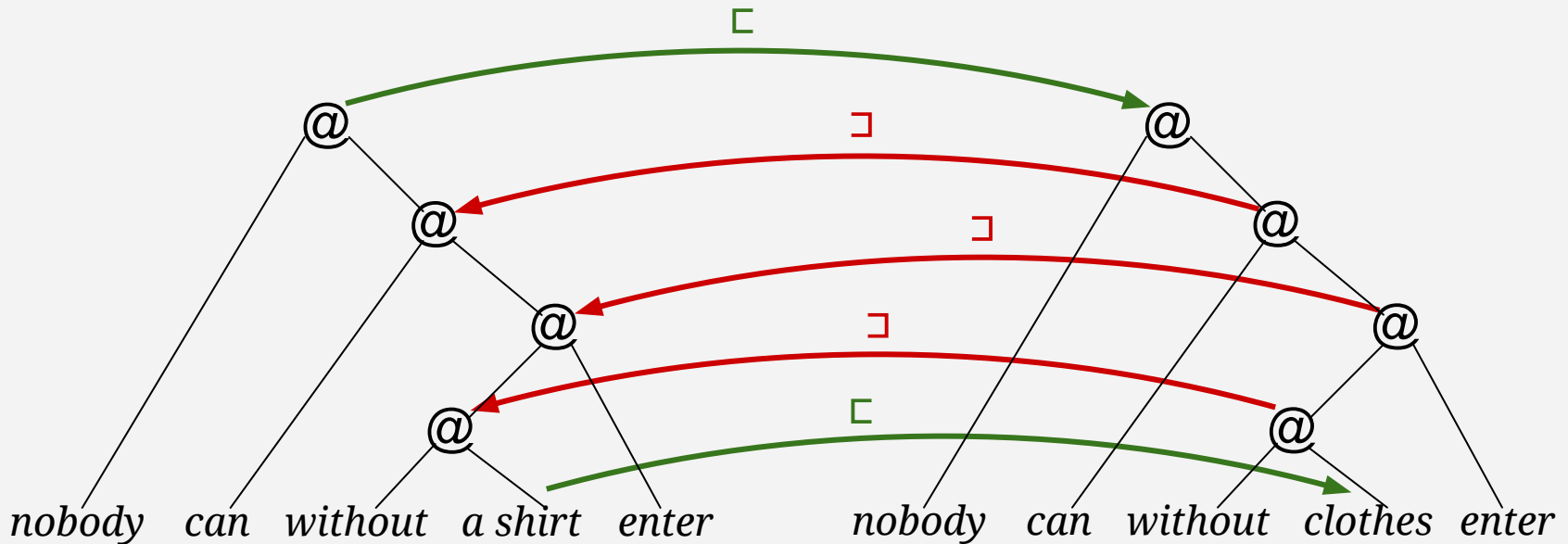
Example: projectivity of *refuse*

	projection	example
	$\equiv \rightarrow \equiv$	
downward monotonicity	$\sqsubset \rightarrow \sqsupset$	<i>refuse to tango</i> \sqsupset <i>refuse to dance</i>
	$\sqsupset \rightarrow \sqsubset$	
	$\# \rightarrow \#$	
switch	$\wedge \rightarrow $	<i>refuse to stay</i> $ $ <i>refuse to go</i>
	$ \rightarrow \#$	<i>refuse to tango</i> $\#$ <i>refuse to waltz</i>
blocks, not swaps	$\smile \rightarrow \#$	

Projecting semantic relations upward

Nobody can enter without a shirt \sqsubset *Nobody can enter without clothes*

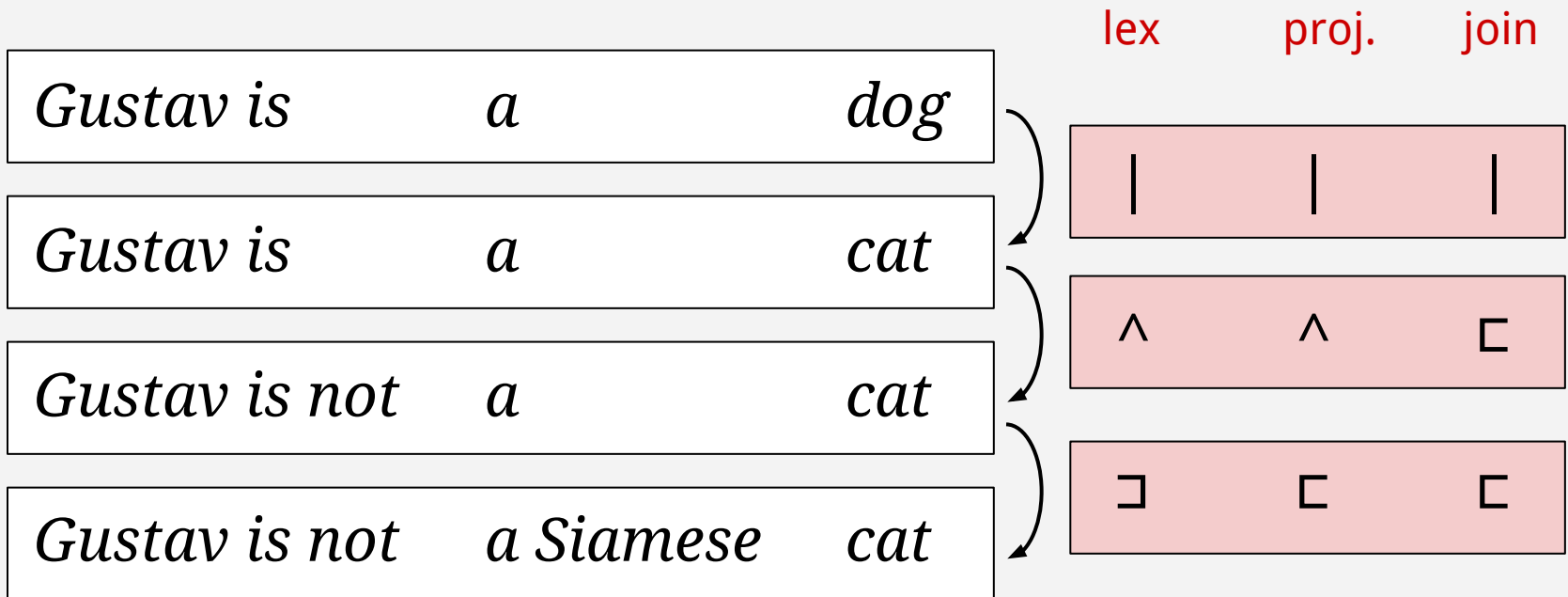
- Assume idealized semantic composition trees
- Propagate lexical semantic relations upward, according to projectivity class of each node on path to root



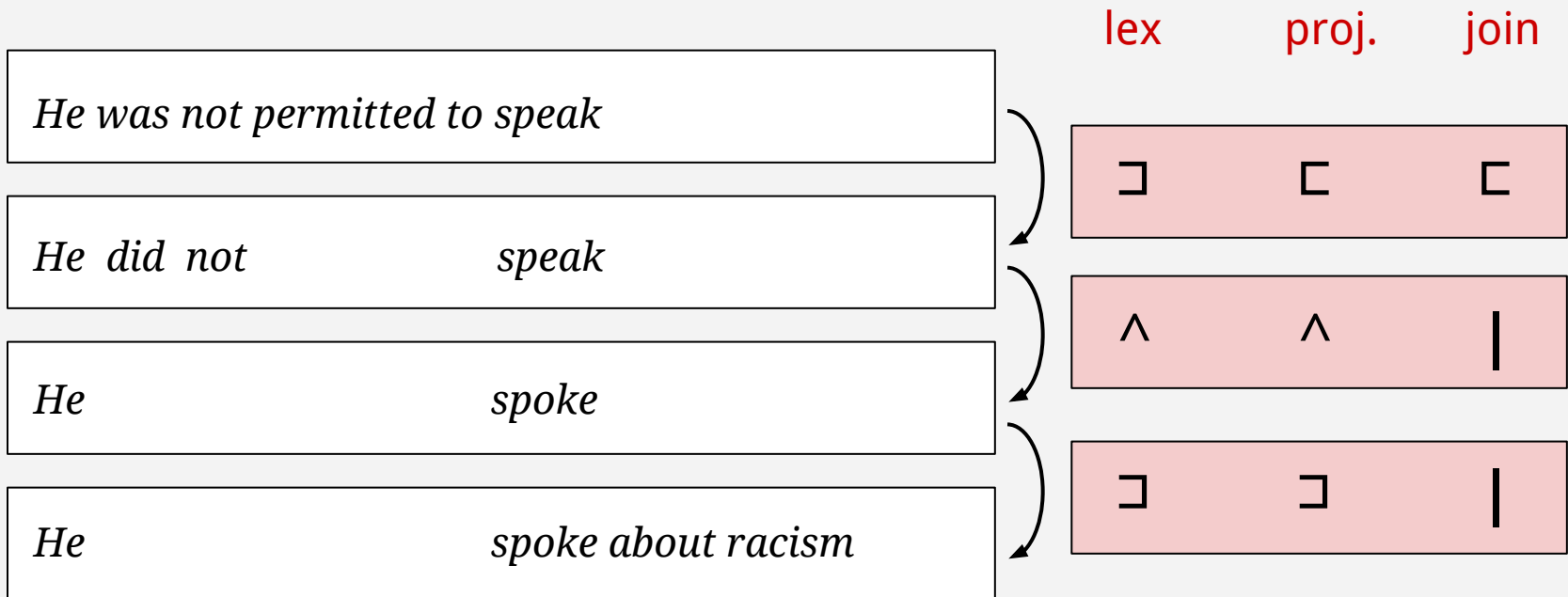
A weak proof procedure

1. Find sequence of edits connecting P and H
 - Insertions, deletions, substitutions, ...
 - E.g., by using a *monolingual aligner* [MacCartney et al. 2008]
2. Determine lexical semantic relation for each edit
 - Substitutions: depends on meaning of substituends: *cat* | *dog*
 - Deletions: \sqsubset by default: *red socks* \sqsubset *socks*
 - But some deletions are special: *not hungry* \wedge *hungry*
 - Insertions are symmetric to deletions: \sqsupset by default
3. Project up to find semantic relation across each edit
4. Join semantic relations across sequence of edits

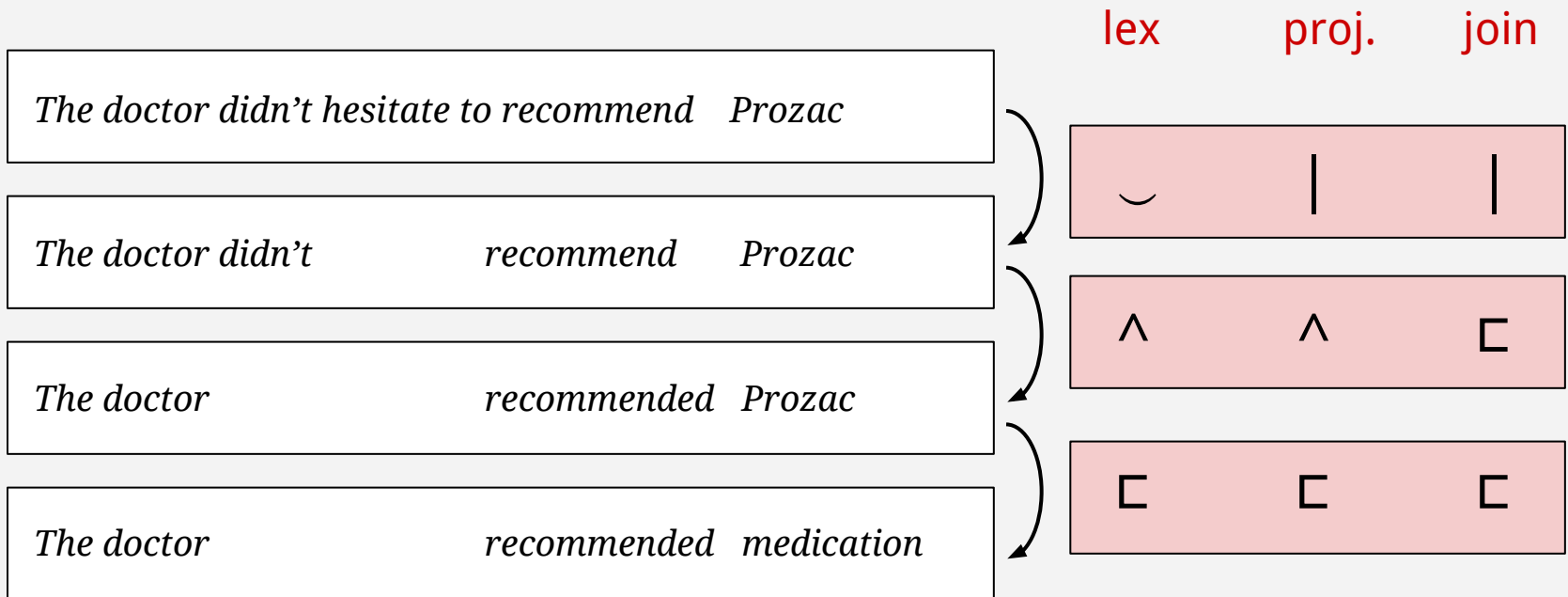
A simple example



An implicative example



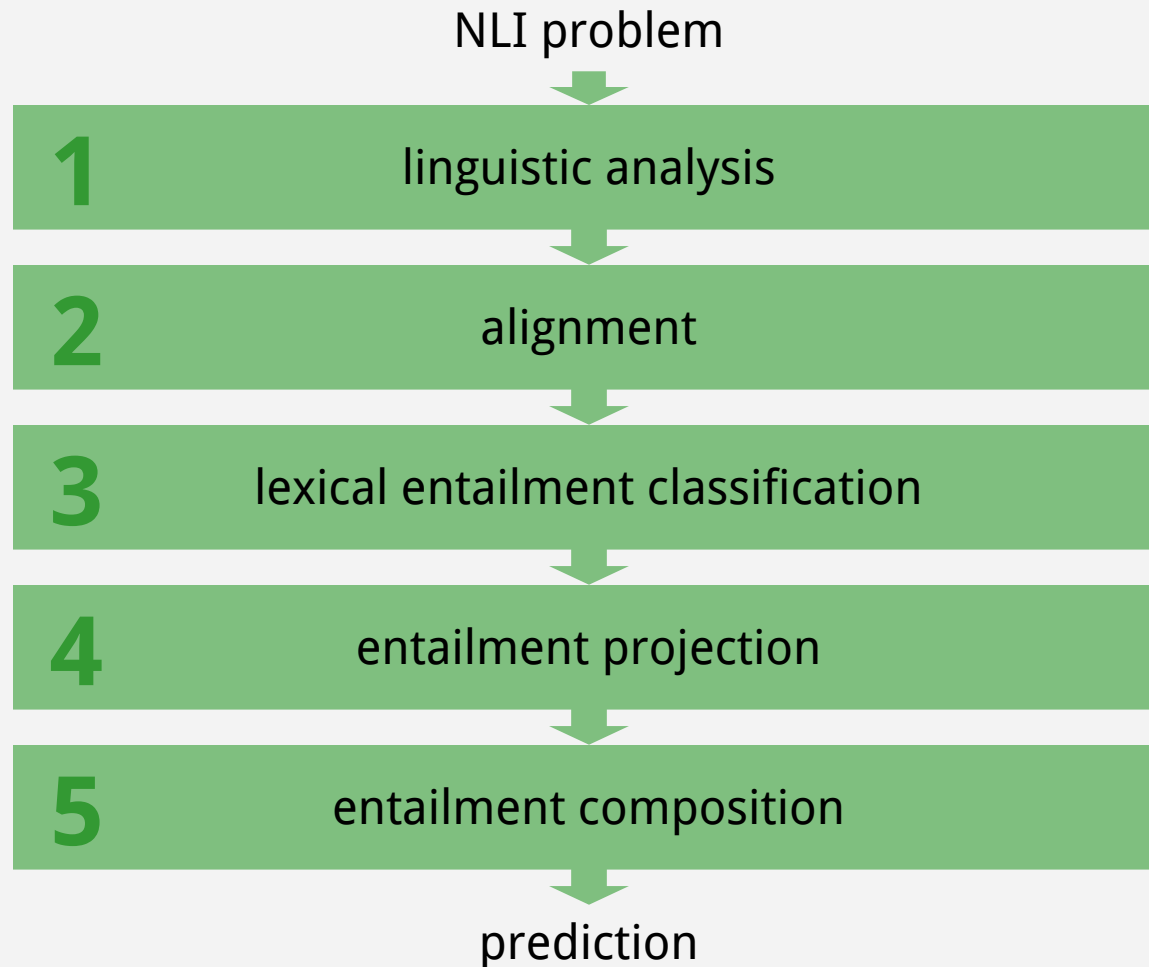
Another implicative example



Outline

- The natural language inference task
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
 - An algebra of semantic relations
 - An account of compositional entailment
 - A weak proof procedure
- **NatLog: implementation & evaluation**
- More recent work by others

The NatLog system



The FraCaS test suite

- 346 “textbook” examples of NLI problems
- 9 sections: quantifiers, plurals, anaphora, ellipsis, ...
- Cons: small size, artificial distribution
- Pros: comprehensive coverage of semantic phenomena

P *No delegate finished the report.*

H *Some delegate finished the report on time.*

no

P *ITEL won more orders than APCOM.*

H *ITEL won some orders.*

yes

P *Smith believed that ITEL had won the contract in 1992.*

H *ITEL won the contract in 1992.*

unk

Key results on FraCaS

- Baseline accuracy: 56% (most common class)
- NatLog accuracy: 70% (32% error reduction)
- Accuracy excl. anaphora, ellipsis, time, verbs: 87%
- Precision over all problems: 90%

The RTE3 test suite

- More “natural” NLI problems; much longer premises
- But not ideal for NatLog
 - Many kinds of inference not addressed by NatLog: paraphrase, temporal reasoning, relation extraction, ...
 - Big edit distance \Rightarrow propagation of errors from atomic model

P *As leaders gather in Argentina ahead of this weekend’s regional talks, Hugo Chávez, Venezuela’s populist president is using an energy windfall to win friends and promote his vision of 21st-century socialism.*

H *Hugo Chávez acts as Venezuela's president.*

yes

Key results on RTE3

system	data	% yes	prec %	rec %	acc %
NatLog	dev	22.5	73.9	32.4	59.3
	test	26.4	70.1	36.1	59.4
Stanford RTE	dev	50.3	68.7	67.0	67.3
	test	50.0	61.8	60.2	60.5
Stanford RTE + NatLog	dev	56.0	69.2	75.2	70.0
	test	54.5	64.5	68.5	64.5

+22 probs
 +36 probs

Outline

- The natural language inference task
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
 - An algebra of semantic relations
 - An account of compositional entailment
 - A weak proof procedure
- NatLog: implementation & evaluation
- **More recent work by others**

Pavlick's dissertation work

- Goal: predict lexical semantic relations for PPDB phrase pairs
- Use much more training data than I did
 - 13,000+ phrase pairs labeled with relations by MTurk
- Use much richer features than I did
 - Including features based on DIRT, PPDB, syntactic paths
- Result: good performance, and a valuable semantic resource

Features	#			≡			□, ⊐			, ^			Other			Average F	
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	Macro	Micro
Majority	52	100	69	–	–	–	–	–	–	–	–	–	–	–	–	14	52
Lexical	59	83	69	35	0	1	15	8	10	5	1	2	19	28	23	21	49
DIRT	64	83	72	45	38	39	13	2	3	–	–	–	14	15	14	26	56
Path	64	58	60	36	13	18	13	12	12	42	16	22	16	45	23	27	41
PPDB	69	81	74	61	53	57	19	2	3	–	–	–	22	27	24	32	60
All	71	80	75	60	52	56	22	10	14	50	27	34	28	30	29	42	60

Angeli & Manning 2014

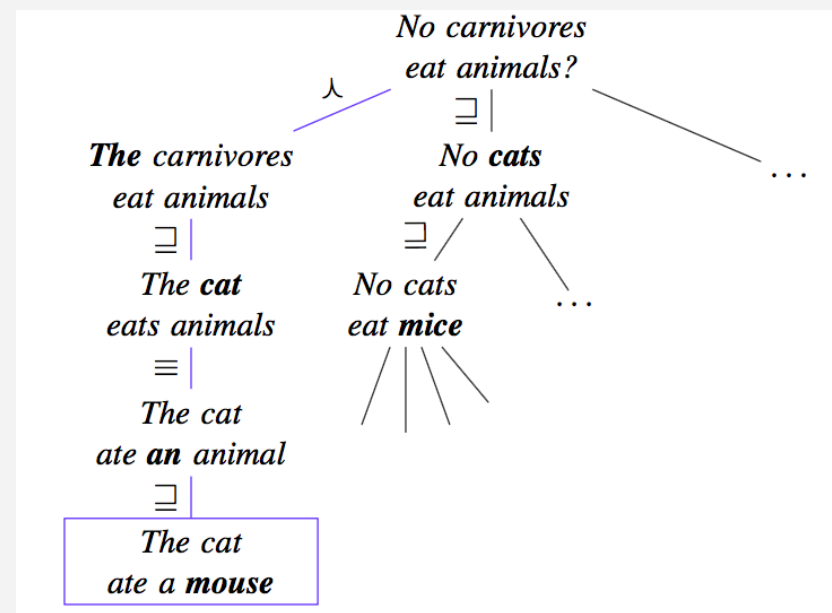
NaturalLI: Natural Logic Inference for Common Sense Reasoning

Can we infer common sense facts from 270M OpenIE facts?

not all birds can fly
noses are used to smell
nobody wants to die
music is used for pleasure

Formulates natural logic as a search problem with costs.

Predicts common sense facts with 49% recall and 91% precision.



Bowman et al. 2014, 2015

- 2014: [Can Recursive Neural Tensor Networks Learn Logical Reasoning?](#)
- 2015: [Recursive Neural Networks Can Learn Logical Semantics](#)

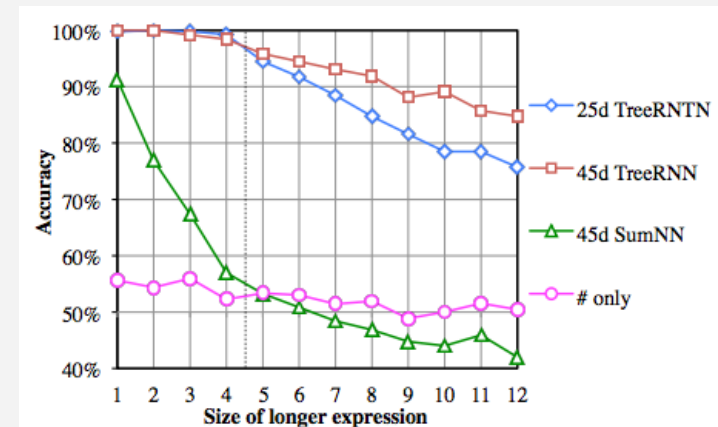
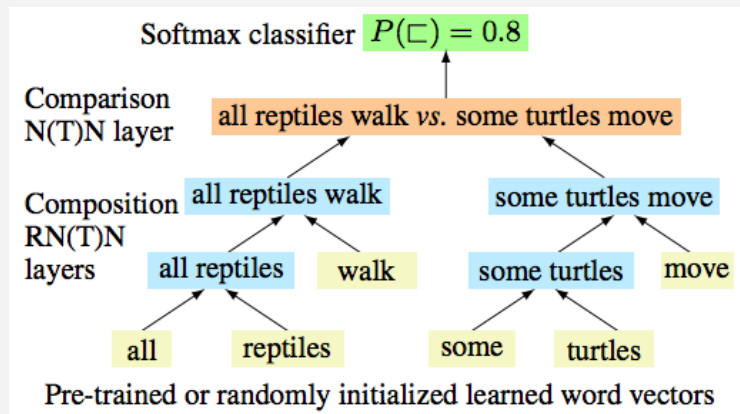


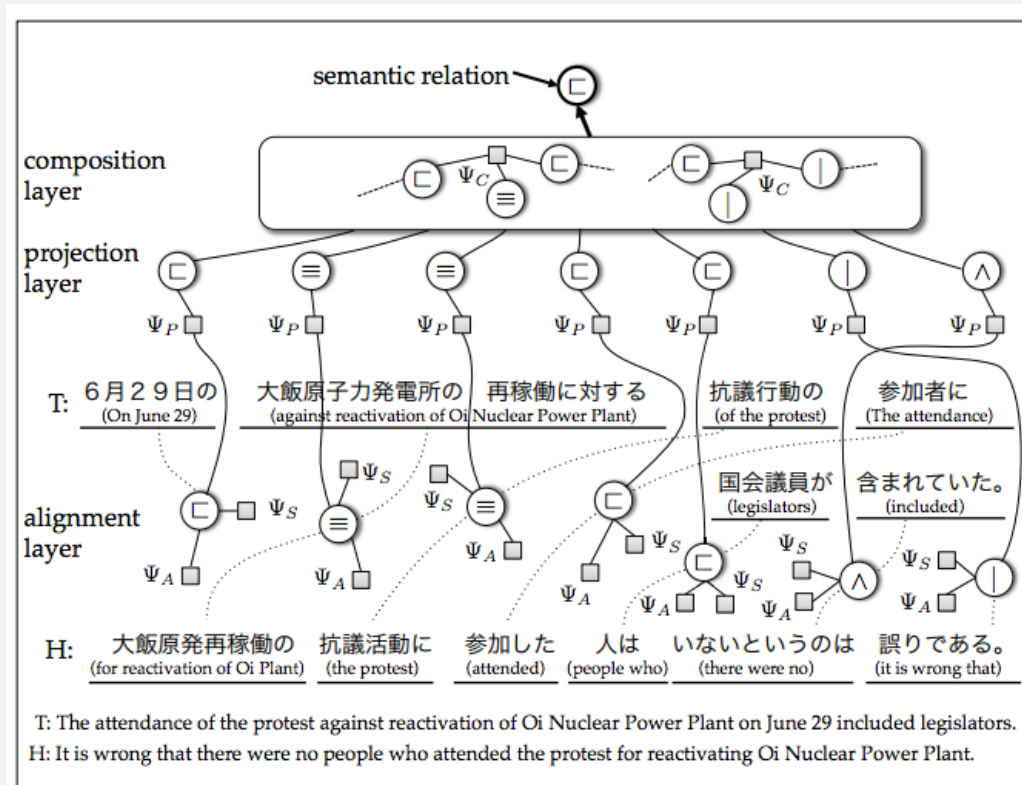
Figure 3: Results on recursive structure. The vertical dotted line marks the size of the longest training examples.

$$(1) \vec{y}_{TreeRNN} = f(\mathbf{M} \begin{bmatrix} \vec{x}^{(l)} \\ \vec{x}^{(r)} \end{bmatrix} + \vec{b})$$

$$(2) \vec{y}_{TreeRNTN} = \vec{y}_{TreeRNN} + f(\vec{x}^{(l)T} \mathbf{T}^{[1\dots n]} \vec{x}^{(r)})$$

Watanabe et al. 2012

A Latent Discriminative Model for Compositional Entailment Relation Recognition Using Natural Logic



Formal underpinnings

Recent work by honest-to-God logicians has helped to secure the theoretical foundations of my approach to natural logic.

- Thomas Icard, III. 2012.
[Inclusion and exclusion in natural language.](#)
- Alex J. Djalali. 2013.
[Synthetic logic.](#)
- Thomas Icard, III and Lawrence Moss. 2014.
[Recent progress on monotonicity.](#)

New dataset: Potts et al.

- Show an image with a caption to an MTurker
- Elicit a novel sentence having a specific relation to caption
- Validate the relation label with other MTurkers
- 140K sentence pairs collected (so far!)

Image caption	Entailment	Contradiction	Independent
Three people with political signs.	People have signs displaying political themes.	Three people have signs promoting their football team.	Men and women are holding up political placards at a rally.
A person working for the city begins cutting down a tree.	A city employee is working outdoors.	The town sheriff is sitting on a tree swing.	A woman who works for the city is using a chainsaw.
A young girl in a white shirt and blue shorts riding high on a swing.	A child is swinging.	A young girl is kneeling in a church.	A child is swinging with friends on equipment in a school playground.

What natural logic can't do

- Not a universal solution for natural language inference
- Many types of inference not amenable to natural logic
 - Paraphrase: *Eve was let go* \equiv *Eve lost her job*
 - Verb/frame alternation: *he drained the oil* \sqsubset *the oil drained*
 - Relation extraction: *Aho, a trader at UBS...* \sqsubset *Aho works for UBS*
 - Common-sense reasoning: *the sink overflowed* \sqsubset *the floor got wet*
 - etc.
- Also, has a weaker proof theory than FOL
 - Can't explain, e.g., de Morgan's laws for quantifiers:
 - *Not all birds fly* \equiv *Some birds don't fly*

What natural logic *can* do

- Natural logic enables precise reasoning about containment, exclusion, and implicativity, while sidestepping the difficulties of translating to logical forms.
- The NatLog system successfully handles a broad range of such inferences, as demonstrated on the FraCaS test suite.
- Ultimately, open-domain natural language inference is likely to require combining disparate reasoners; natural logic is a good candidate to be a component of such a system.



Thanks! Questions?

Backup slides



Backup slides follow

Some simple inferences

No state completely forbids casino gambling.

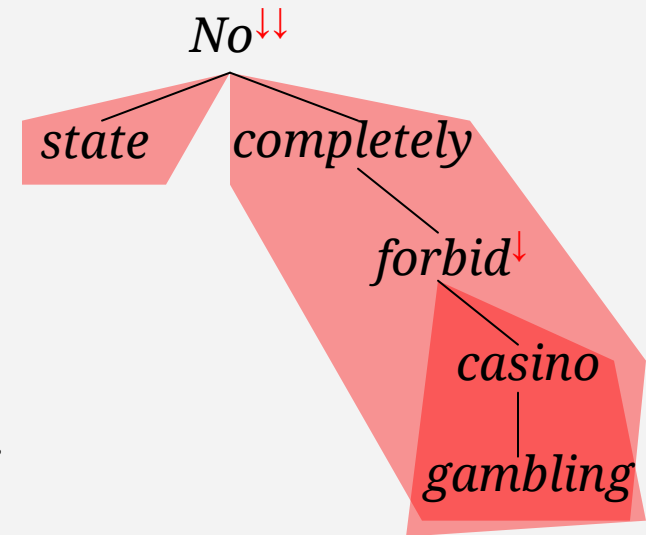
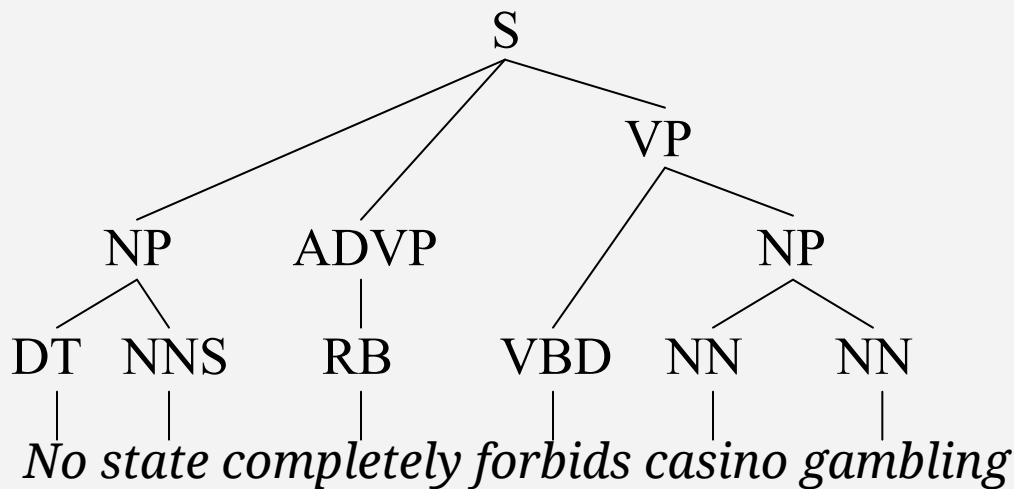
OK *No western state completely forbids casino gambling.*
No state completely forbids [red box] gambling.
Few or no states completely forbid casino gambling.

No *No state completely forbids casino gambling for kids.*
No state restricts gambling.
No state or city completely forbids casino gambling.

What kind of NLI system could predict this?

Step 1: Linguistic analysis

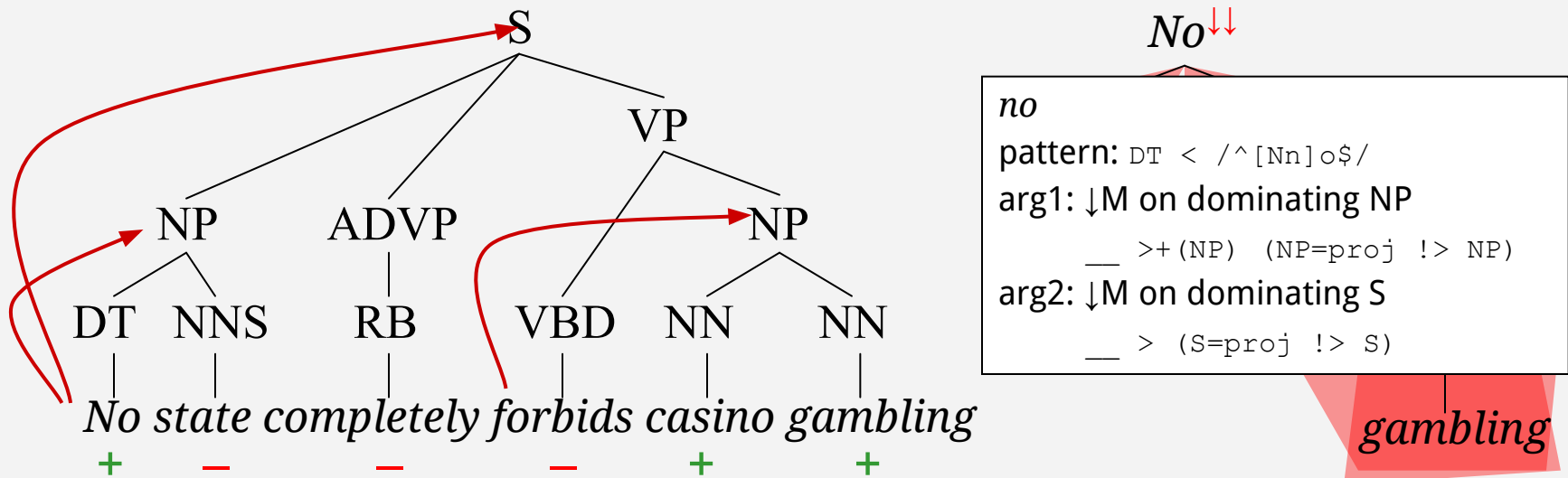
- Tokenize & parse input sentences
- Identify items w/ special projectivity & determine scope
- Problem: PTB-style parse tree \neq semantic structure!



- Solution: specify scope in PTB trees using Tregex [Levy & Andrew 06]

Step 1: Linguistic analysis

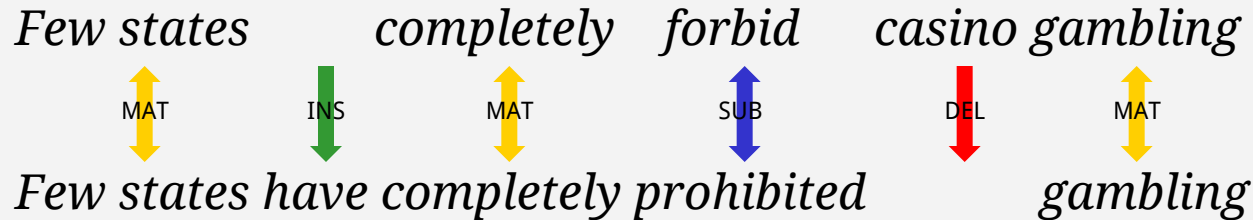
- Tokenize & parse input sentences
- Identify items w/ special projectivity & determine scope
- Problem: PTB-style parse tree \neq semantic structure!



- Solution: specify scope in PTB trees using Tregex [Levy & Andrew 06]

Step 2: Alignment

- Phrase-based alignments: symmetric, many-to-many
- Can view as sequence of *atomic edits*: DEL, INS, SUB, MAT



- Ordering of edits defines path through intermediate forms
 - Need not correspond to sentence order
- Decomposes problem into atomic entailment problems
- (I proposed an alignment system in an EMNLP-08 paper)

Running example

<i>P</i>	<i>Jimmy Dean</i>	<i>refused to</i>			<i>move</i>	<i>without</i>	<i>blue</i>	<i>jeans</i>
<i>H</i>	<i>James Dean</i>		<i>did</i>	<i>n't</i>	<i>dance</i>	<i>without</i>		<i>pants</i>
<i>edit index</i>	1	2	3	4	5	6	7	8
<i>edit type</i>	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB

OK, the example is contrived, but it compactly exhibits containment, exclusion, and implicativity

Step 3: Lexical entailment classification

- Predict basic semantic relation for each edit, based solely on lexical features, independent of context
- Feature representation:
 - WordNet features: synonymy, hyponymy, antonymy
 - Other relatedness features: Jiang-Conrath (WN-based), NomBank
 - String and lemma similarity, based on Levenshtein edit distance
 - Lexical category features: *prep*, *poss*, *art*, *aux*, *pron*, *pn*, etc.
 - Quantifier category features
 - Implication signatures (for DEL edits only)
- Decision tree classifier
 - Trained on 2,449 hand-annotated lexical entailment problems
 - Very low training error — captures relevant distinctions

Running example

<i>P</i>	<i>Jimmy Dean</i>	<i>refused to</i>			<i>move</i>	<i>without</i>	<i>blue</i>	<i>jeans</i>
<i>H</i>	<i>James Dean</i>		<i>did</i>	<i>n't</i>	<i>dance</i>	<i>without</i>		<i>pants</i>
<i>edit index</i>	1	2	3	4	5	6	7	8
<i>edit type</i>	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
<i>lex feats</i>	strsim= 0.67	implic: +/o	cat:aux	cat:neg	hypo			hyper
<i>lex entrel</i>	≡		≡	^	⊐	≡	⊐	⊐

Step 4: entailment projection

<i>P</i>	<i>Jimmy Dean</i>	<i>refused to</i>			<i>move</i>	<i>without</i>	<i>blue</i>	<i>jeans</i>
<i>H</i>	<i>James Dean</i>		<i>did</i>	<i>n't</i>	<i>dance</i>	<i>without</i>		<i>pants</i>
<i>edit index</i>	1	2	3	4	5	6	7	8
<i>edit type</i>	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
<i>lex feats</i>	strsim=0.67	implic:+/o	cat:aux	cat:neg	hypo			hyper
<i>lex entrel</i>	≡		≡	^	⊐	≡	⊐	⊐
<i>projectivity</i>	↑	↑	↑	↑	↓	↓	↑	↑
<i>atomic entrel</i>	≡		≡	^	⊐	≡	⊐	⊐

inversion

Step 5: Entailment composition

<i>P</i>	<i>Jimmy Dean</i>	<i>refused to</i>			<i>move</i>	<i>without</i>	<i>blue</i>	<i>jeans</i>
<i>H</i>	<i>James Dean</i>		<i>did</i>	<i>n't</i>	<i>dance</i>	<i>without</i>		<i>pants</i>
<i>edit index</i>	1	2	3	4	5	6	7	8
<i>edit type</i>	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
<i>lex feats</i>	strsim=0.67	implic:+/o	cat:aux	cat:neg	hypo			hyper
<i>lex entrel</i>	≡		≡	^	⊐	≡	⊐	⊐
<i>projectivity</i>	↑	↑	↑	↑	↓	↓	↑	↑
<i>atomic entrel</i>	≡		≡	^	⊐	≡	⊐	⊐
<i>composition</i>	≡			⊐	⊐	⊐	⊐	⊐

interesting

final answer

The FraCaS test suite

- FraCaS: mid-90s project in computational semantics
- 346 “textbook” examples of NLI problems
 - examples on next slide
- 9 sections: quantifiers, plurals, anaphora, ellipsis, ...
- 3 possible answers: *yes*, *no*, *unknown* (not balanced!)
- 55% single-premise, 45% multi-premise (excluded)


FraCaS examples

- P *No delegate finished the report.*
- H *Some delegate finished the report on time.* no
- P *At most ten commissioners spend time at home.*
- H *At most ten commissioners spend a lot of time at home.* yes
- P *Either Smith, Jones or Anderson signed the contract.*
- H *Jones signed the contract.* unk
- P *Dumbo is a large animal.*
- H *Dumbo is a small animal.* no
- P *ITEL won more orders than APCOM.*
- H *ITEL won some orders.* yes
- P *Smith believed that ITEL had won the contract in 1992.*
- H *ITEL won the contract in 1992.* unk

Results on FraCaS

System	#	prec %	rec %	acc %
most common class	183	55.7	100.0	55.7
MacCartney & M. 07	183	68.9	60.8	59.6
MacCartney & M. 08	183	89.3	65.7	70.5

27% error reduction

A black arrow originates from the text "27% error reduction" and points to the value "70.5" in the accuracy column of the table, which is highlighted with a red oval.

Results on FraCaS

System	#	prec %	rec %	acc %
most common class	183	55.7	100.0	55.7
MacCartney & M. 07	183	68.9	60.8	59.6
MacCartney & M. 08	183	89.3	65.7	70.5

27% error reduction

§	Category	#	prec %	rec %	acc %
1	Quantifiers	44	95.2	100.0	97.7
2	Plurals	24	90.0	64.3	75.0
3	Anaphora	6	100.0	60.0	50.0
4	Ellipsis	25	100.0	5.3	24.0
5	Adjectives	15	71.4	83.3	80.0
6	Comparatives	16	88.9	88.9	81.3
7	Temporal	36	85.7	70.6	58.3
8	Verbs	8	80.0	66.7	62.5
9	Attitudes	9	100.0	83.3	88.9
1, 2, 5, 6, 9		108	90.4	85.5	87.0

in largest category,
all but one correct

high accuracy
in sections
most amenable
to natural logic

high precision
even outside
areas of expertise

FraCaS confusion matrix

		guess			total
		<i>yes</i>	<i>no</i>	<i>unk</i>	
gold	<i>yes</i>	67	4	31	102
	<i>no</i>	1	16	4	21
	<i>unk</i>	7	7	46	60
total		75	27	81	183

The RTE3 test suite

- RTE: more “natural” natural language inference problems
- Much longer premises: average 35 words (vs. 11)
- Binary classification: *yes* and *no*
- RTE problems not ideal for NatLog
 - Many kinds of inference not addressed by NatLog: paraphrase, temporal reasoning, relation extraction, ...
 - Big edit distance \Rightarrow propagation of errors from atomic model

RTE3 examples

P *As leaders gather in Argentina ahead of this weekend's regional talks, Hugo Chávez, Venezuela's populist president is using an energy windfall to win friends and promote his vision of 21st-century socialism.*

H *Hugo Chávez acts as Venezuela's president.*

yes

P *Democrat members of the Ways and Means Committee, where tax bills are written and advanced, do not have strong small business voting records.*

H *Democrat members had strong small business voting records.*

no

(These examples are probably easier than average for RTE.)

Results on RTE3 data

system	data	% yes	prec %	rec %	acc %
RTE3 best (LCC)	test				80.0
RTE3 2nd best (LCC)	test				72.2
RTE3 average other 24	test				60.5
NatLog	dev	22.5	73.9	32.3	59.3
	test	26.4	70.1	36.1	59.4

(each data set contains 800 problems)

- Accuracy is unimpressive, but precision is relatively high
- Maybe we can achieve high precision on a subset?
- Strategy: hybridize with broad-coverage RTE system
 - As in Bos & Markert 2006

A simple bag-of-words model

P \ H	H		
	<i>Dogs</i>	<i>hate</i>	<i>figs</i>
<i>Dogs</i>	1.00	0.00	0.33
<i>do</i>	0.67	0.00	0.00
<i>n't</i>	0.33	0.25	0.00
<i>like</i>	0.00	0.25	0.25
<i>fruit</i>	0.00	0.00	0.40
max	1.00	0.25	0.40
IDF	0.43	0.55	0.80
$P(h P)$	1.00	0.47	0.48
$P(H P)$	0.23		

similarity scores on [0, 1]
for each pair of words
(I used a really simple-minded
similarity function based on
Levenshtein string-edit distance)

max sim for each hyp word

how rare each word is

$= (\max \text{ sim})^{\text{IDF}}$

$= \prod_h P(h | P)$

A simple bag-of-words model

P \ H	<i>Dogs</i>	<i>hate</i>	<i>figs</i>	max	IDF	P(p H)	P(P H)
<i>Dogs</i>	1.00	0.00	0.33	1.00	0.43	1.00	
<i>do</i>	0.67	0.00	0.00	0.67	0.11	0.96	
<i>n't</i>	0.33	0.25	0.00	0.33	0.05	0.95	0.43
<i>like</i>	0.00	0.25	0.25	0.25	0.25	0.71	
<i>fruit</i>	0.00	0.00	0.40	0.40	0.46	0.66	
max	1.00	0.25	0.40	max sim for each hyp word			
IDF	0.43	0.55	0.80	how rare each word is			
P(h P)	1.00	0.47	0.48	= (max sim)^IDF			
P(H P)		0.23		= $\prod_h P(h P)$			

Results on RTE3 data

system	data	% yes	prec %	rec %	acc %
RTE3 best (LCC)	test				80.0
RTE3 2nd best (LCC)	test				72.2
RTE3 average other 24	test				60.5
NatLog	dev	22.5	73.9	32.3	59.3
	test	26.4	70.1	36.1	59.4
BoW (bag of words)	dev	50.6	70.1	68.9	68.9
	test	51.2	62.4	70.0	63.0

+20 probs

(each data set contains 800 problems)

Combining BoW and NatLog

- MaxEnt classifier
- BoW features: $P(H | P)$, $P(P | H)$
- NatLog features:
7 boolean features encoding predicted semantic relation

Results on RTE3 data

system	data	% yes	prec %	rec %	acc %
RTE3 best (LCC)	test				80.0
RTE3 2nd best (LCC)	test				72.2
RTE3 average other 24	test				60.5
NatLog	dev	22.5	73.9	32.3	59.3
	test	26.4	70.1	36.1	59.4
BoW (bag of words)	dev	50.6	70.1	68.9	68.9
	test	51.2	62.4	70.0	63.0
BoW + NatLog	dev	50.7	71.4	70.4	70.3
	test	56.1	63.0	69.0	63.4

+11 probs
 +3 probs

(each data set contains 800 problems)

Problem: NatLog is *too* precise?

- Error analysis reveals a characteristic pattern of mistakes:
 - Correct answer is *yes*
 - Number of edits is large (>5) (this is typical for RTE)
 - NatLog predicts \sqsubset or \equiv for all but one or two edits
 - But NatLog predicts some other relation for remaining edits!
 - Most commonly, it predicts \supset for an insertion (e.g., “acts as”)
 - Result of relation composition is thus $\#$, i.e. *no*
- Idea: make it more forgiving, by adding features
 - Number of edits
 - Proportion of edits for which predicted relation is not \sqsubset or \equiv

Results on RTE3 data

system	data	% yes	prec %	rec %	acc %	
RTE3 best (LCC)	test				80.0	
RTE3 2nd best (LCC)	test				72.2	
RTE3 average other 24	test				60.5	
NatLog	dev	22.5	73.9	32.3	59.3	
	test	26.4	70.1	36.1	59.4	
BoW (bag of words)	dev	50.6	70.1	68.9	68.9	
	test	51.2	62.4	70.0	63.0	
BoW + NatLog	dev	50.7	71.4	70.4	70.3	+13 probs
	test	56.1	63.0	69.0	63.4	+8 probs
BoW + NatLog + other	dev	52.7	70.9	72.6	70.5	
	test	58.7	63.0	72.2	64.0	