Natural logic and textual inference

Bill MacCartney

CS224U

12 May 2014

Textual inference 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

The textual inference task

- 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 textual 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?
  - Textual inference: P $\rightarrow$ Q? Paraphrase: P $\leftrightarrow$ Q?

Textual inference 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.*

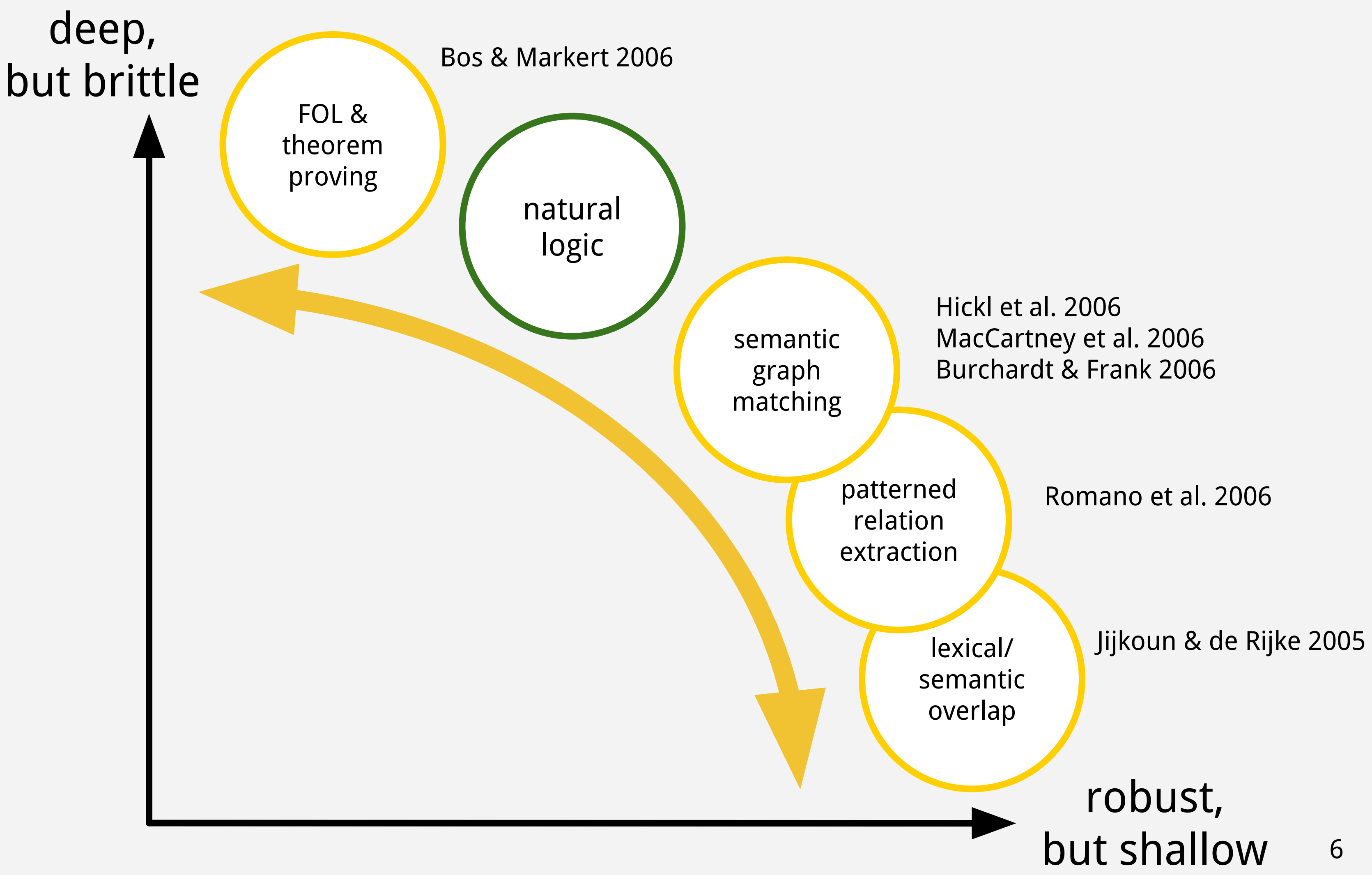
- If you can't recognize that P implies H, then you haven't really understood P (or H)

- Thus, a capacity for textual 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

Approaches to textual inference



Outline

- The textual inference task
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

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 ⊏ *tall girl* ⊏ *girl*
- Generalizing instances or classes into superclasses:

  Einstein ⊏ *a physicist* ⊏ *a scientist*
- Spatial & temporal broadening:

  in Palo Alto ⊏ *in California*, *this month* ⊏ *this year*
- Relaxing modals: *must* ⊏ *could*, *definitely* ⊏ *probably* ⊏ *maybe*
- Relaxing quantifiers: *six* ⊏ *several* ⊏ *some*
- Dropping conjuncts, adding disjuncts:

  danced and sang ⊏ *sang* ⊏ *hummed or sang*

Downward monotonicity

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

  My cat did not eat a rat ⇐ My cat did not eat a rodent

- But also many other negative or restrictive expressions!

  No cats ate rats ⇐ No cats ate rodents
  Every rat fears my cat ⇐ Every rodent fears my cat
  My cat ate at most three rats ⇐ My cat ate at most three rodents
  If my cat eats a rat, he'll puke ⇐ If my cat eats a rodent, he'll puke
  My cat avoids eating rats ⇐ My cat avoids eating rodents
  My cat denies eating a rat ⇐ My cat denies eating a rodent
  My cat rarely eats rats ⇐ 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 $\sqsubset$ *rodent*, *eat* $\sqsubset$ *consume*, *this morning* $\sqsubset$ *today*, *most* $\sqsubset$ *some*
- Monotonicity classes for semantic functions
  - Upward monotone: *some rats dream* $\sqsubset$ *some rodents dream*
  - Downward monotone: *no rats dream* $\sqsupset$ *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 $\sqsubset$ *Gustav is not a Siamese cat*

Outline

- Introduction
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

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*

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 problem of textual inference

Gustav is	*a*	*dog*
Gustav is	*a*	*cat*
Gustav is not	*a*	*cat*
Gustav is not	*a Siamese*	*cat*

$\mid$	alternation
$^$	negation
$\sqsubset$	forward entailment

$\sqsubset$	forward entailment

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

Some simple inferences

No state completely forbids casino gambling.

OK *No western state completely forbids casino gambling.*
No state completely forbids 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.

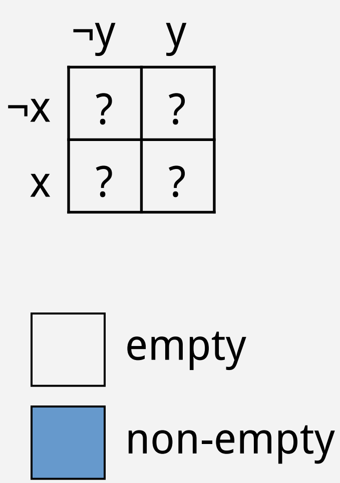
What kind of textual inference system could predict this?

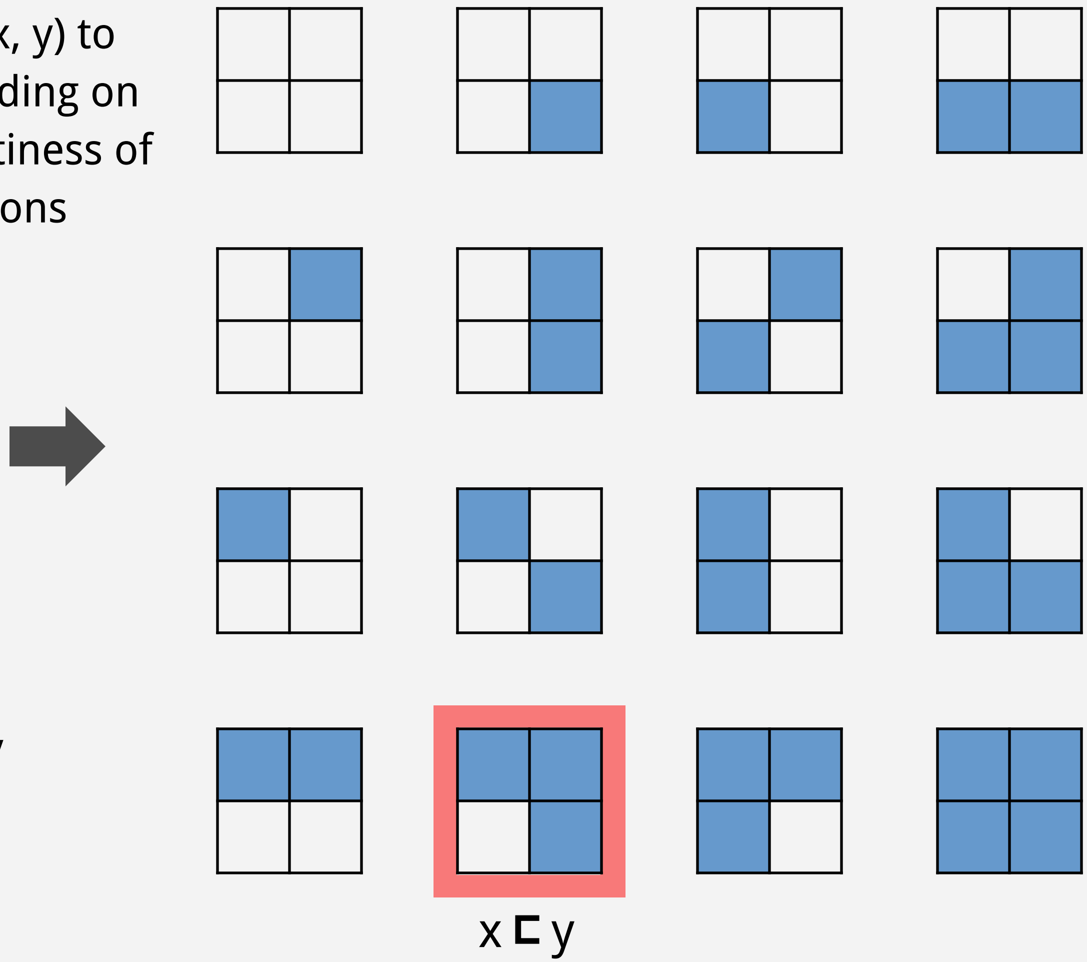
Semantic relations in past work

	X is a couch / *X is a sofa*	*X is a crow* / *X is a bird*	*X is a fish* / *X is a carp*	*X is a hippo* / *X is hungry*	*X is a man* / *X is a woman*
2-way RTE1,2,3	Yes entailment		No non-entailment		
3-way RTE4, FraCaS, PARC	Yes entailment		Unknown non-entailment		No contradiction
containment Sánchez-Valencia	$P \equiv Q$ equivalence	$P \sqsubset Q$ forward entailment	$P \sqsupset Q$ 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



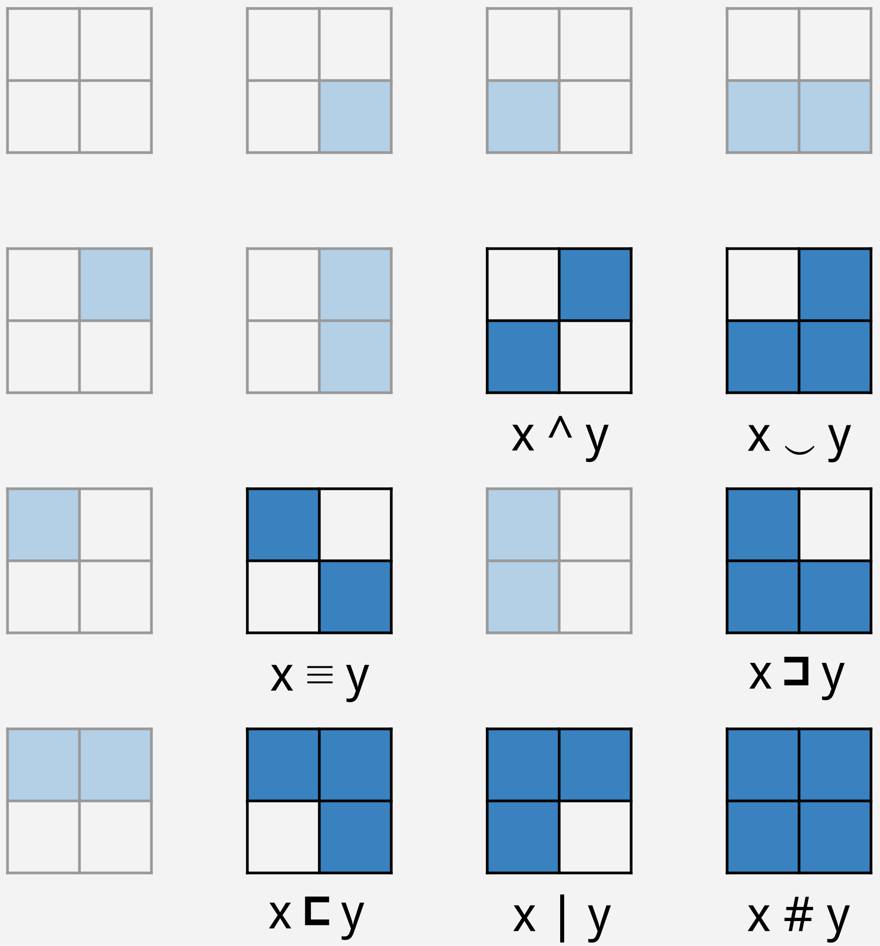


16 elementary set relations

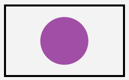
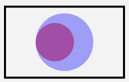
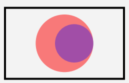
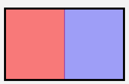
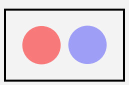
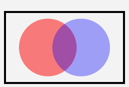
But 9 of 16 are degenerate: either *x* or *y* is either empty or universal.

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

We therefore focus on the remaining seven relations.



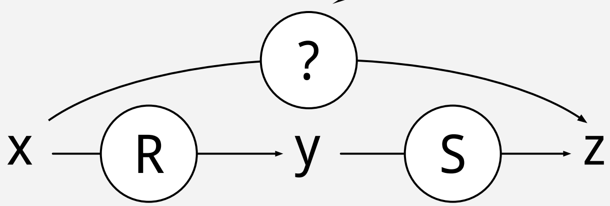
7 basic semantic relations

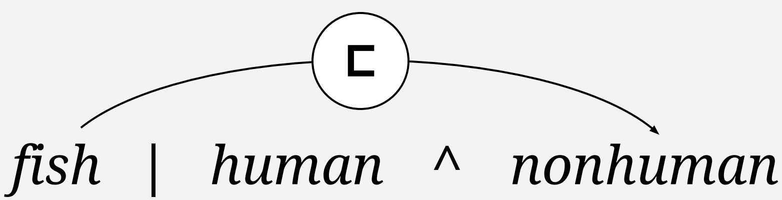
	$x \equiv y$	equivalence	*couch* ≡ *sofa*
	$x \sqsubset y$	forward entailment (strict)	*crow* ⊏ *bird*
	$x \sqsupset y$	reverse entailment (strict)	*European* ⊐ *French*
	$x \wedge y$	negation (exhaustive exclusion)	*human* ^ *nonhuman*
	$x \mid y$	alternation (non-exhaustive exclusion)	*cat* | *dog*
	$x \smile y$	cover (exhaustive non-exclusion)	*animal* ‿ *nonhuman*
	$x \# y$	independence	*hungry* # *hippo*

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

Joining semantic relations

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





$\equiv$	$\bowtie$	$\equiv$	$\Rightarrow$	$\equiv$
$\sqsubset$	$\bowtie$	$\sqsubset$	$\Rightarrow$	$\sqsubset$
$\sqsupset$	$\bowtie$	$\sqsupset$	$\Rightarrow$	$\sqsupset$
^	$\bowtie$	^	$\Rightarrow$	$\equiv$
R	$\bowtie$	$\equiv$	$\Rightarrow$	R
$\equiv$	$\bowtie$	R	$\Rightarrow$	R

Some joins yield unions of relations

What is $| \bowtie |$?

$x \mid y$	$y \mid z$	$x \; ? \; z$
couch | *table*	*table* | *sofa*	*couch* ≡ *sofa*
pistol | *knife*	*knife* | *gun*	*pistol* ⊏ *gun*
dog | *cat*	*cat* | *terrier*	*dog* ⊐ *terrier*
rose | *orchid*	*orchid* | *daisy*	*rose* | *daisy*
woman | *frog*	*frog* | *Eskimo*	*woman* # *Eskimo*

$$| \bowtie | \quad \Rightarrow \quad \bigcup \{\equiv, \sqsubset, \sqsupset, \mid, \#\}$$

The complete join table

$\bowtie$	$\equiv$	$\sqsubset$	$\sqsupset$	$^$	$\mid$	$\smile$	$\#$
$\equiv$	$\equiv$	$\sqsubset$	$\sqsupset$	$^$	$\mid$	$\smile$	$\#$
$\sqsubset$	$\sqsubset$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\mid$	$\mid$	$\sqsubset^\mid\smile\#$	$\sqsubset\mid\#$
$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\smile\#$	$\sqsupset$	$\smile$	$\sqsupset^\mid\smile\#$	$\smile$	$\sqsupset\smile\#$
$^$	$^$	$\smile$	$\mid$	$\equiv$	$\sqsupset$	$\sqsubset$	$\#$
$\mid$	$\mid$	$\sqsubset^\mid\smile\#$	$\mid$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\sqsubset$	$\sqsubset\mid\#$
$\smile$	$\smile$	$\smile$	$\sqsupset^\mid\smile\#$	$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\smile\#$	$\sqsupset\smile\#$
$\#$	$\#$	$\sqsubset\smile\#$	$\sqsupset\mid\#$	$\#$	$\sqsupset\mid\#$	$\sqsubset\smile\#$	$\equiv\sqsubset\sqsupset^\mid\smile\#$

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 (= monoticity++)

- 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 $\equiv$, $\sqsubset$, $\sqsupset$, #)
- But what about the other relations (^, |, $\smile$)?
- We'll categorize semantic functions based on how they *project* the basic semantic relations

Example: projectivity of *not*

projection			example
$\equiv$	$\rightarrow$	$\equiv$	*not happy* $\equiv$ *not glad*
$\sqsubset$	$\rightarrow$	$\sqsupset$	*didn't kiss* $\sqsupset$ *didn't touch*
$\sqsupset$	$\rightarrow$	$\sqsubset$	*isn't European* $\sqsubset$ *isn't French*
#	$\rightarrow$	#	*isn't swimming* # *isn't hungry*
$\wedge$	$\rightarrow$	$\wedge$	*not human* $\wedge$ *not nonhuman*
$\mid$	$\rightarrow$	$\smile$	*not French* $\smile$ *not German*
$\smile$	$\rightarrow$	$\mid$	*not more than 4* $\mid$ *not less than 6*

downward monotonicity

swaps these too

Example: projectivity of *refuse*

projection	example
$\equiv \;\rightarrow\; \equiv$	
$\sqsubset \;\rightarrow\; \sqsupset$	*refuse to tango* $\sqsupset$ *refuse to dance*
$\sqsupset \;\rightarrow\; \sqsubset$	
$\# \;\rightarrow\; \#$	
$\wedge \;\rightarrow\; \mid$	*refuse to stay* $\mid$ *refuse to go*
$\mid \;\rightarrow\; \#$	*refuse to tango* $\#$ *refuse to waltz*
$\smile \;\rightarrow\; \#$	

downward monotonicity

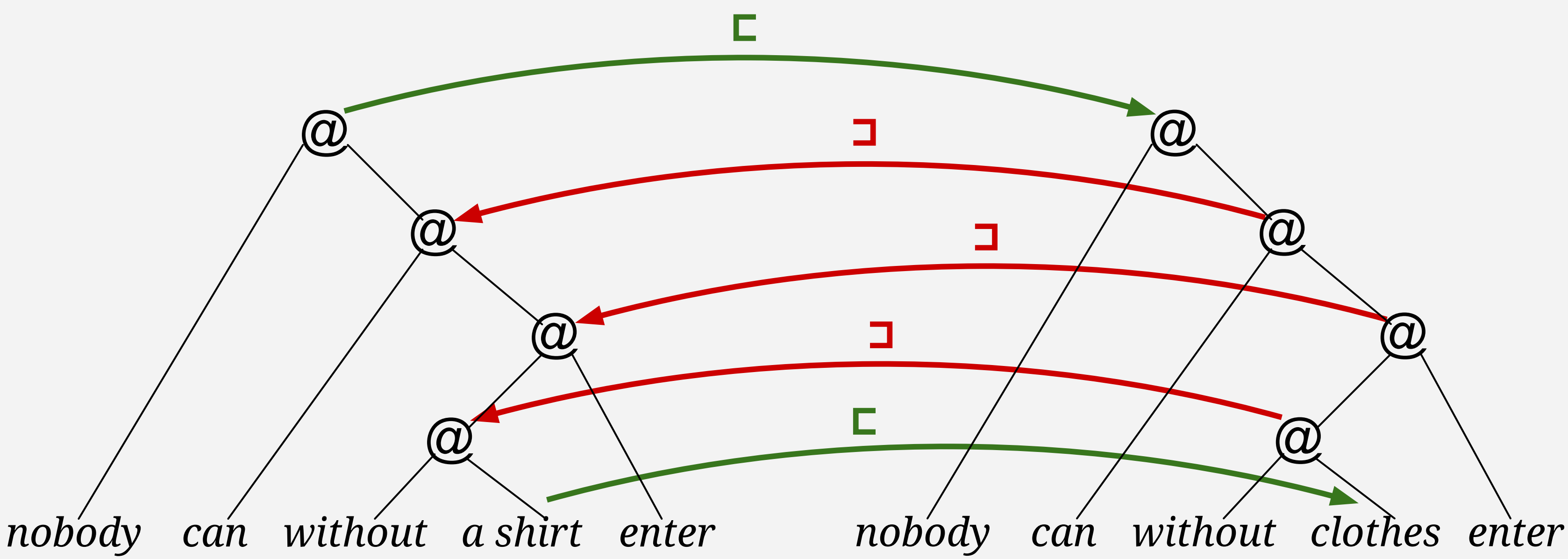
switch

blocks, not swaps

Projecting semantic relations upward

Nobody can enter without a shirt ⊏ *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: ⊏ by default: *red socks* ⊏ *socks*
   - But some deletions are special: *not hungry* ^ *hungry*
   - Insertions are symmetric to deletions: ⊐ by default

3. Project up to find semantic relation across each edit

4. Join semantic relations across sequence of edits

A simple example

Gustav is a dog

Gustav is a cat

Gustav is not a cat

Gustav is not a Siamese cat

lex	proj.	join
$\mid$	$\mid$	$\mid$
$\wedge$	$\wedge$	$\sqsubset$
$\sqsupset$	$\sqsubset$	$\sqsubset$

Outline

- Introduction
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

The NatLog system

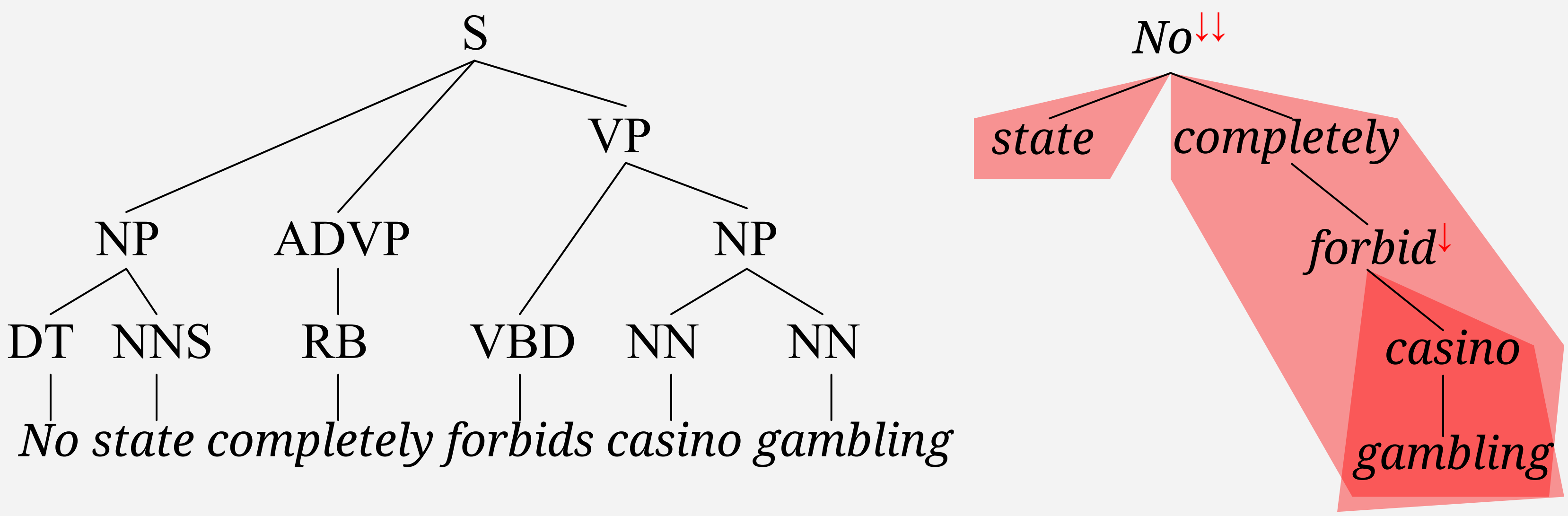
textual inference problem

1. linguistic analysis
2. alignment
3. lexical entailment classification
4. entailment projection
5. entailment composition

prediction

Step 1: Linguistic analysis

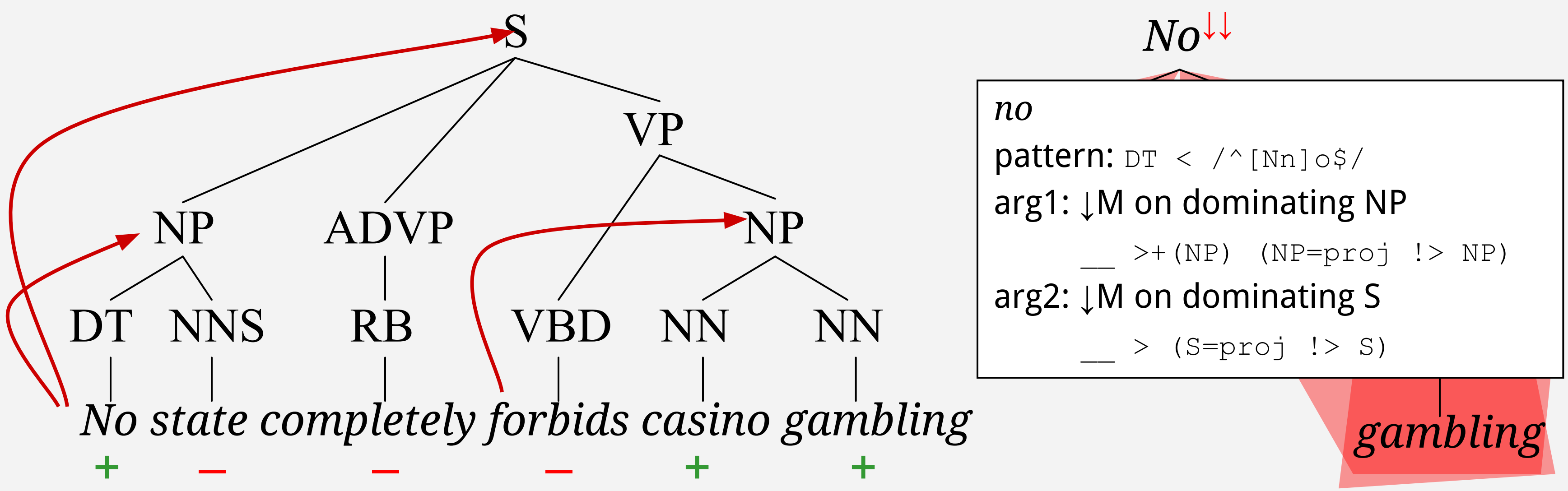
- Tokenize & parse input sentences
- Identify items w/ special projectivity & determine scope
- Problem: PTB-style parse tree ≠ semantic structure!



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

Step 1: Linguistic analysis

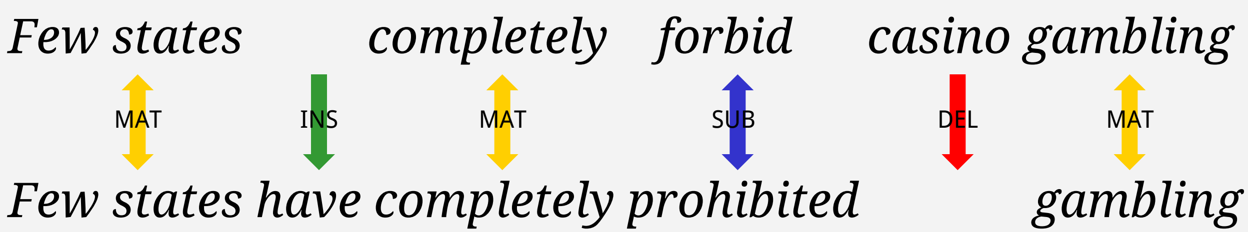
- Tokenize & parse input sentences
- Identify items w/ special projectivity & determine scope
- Problem: PTB-style parse tree ≠ semantic structure!



- 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

P	*Jimmy Dean*	*refused to*			*move*	*without*	*blue*	*jeans*
H	*James Dean*		*did*	*n't*	*dance*	*without*		*pants*
edit index	1	2	3	4	5	6	7	8
edit type	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

P	*Jimmy Dean*	*refused to*			*move*	*without*	*blue*	*jeans*
H	*James Dean*		*did*	*n't*	*dance*	*without*		*pants*
edit index	1	2	3	4	5	6	7	8
edit type	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
lex feats	strsim= 0.67	implic: +/o	cat:aux	cat:neg	hypo			hyper
lex entrel	$\equiv$	$\mid$	$\equiv$	$\wedge$	$\sqsupset$	$\equiv$	$\sqsubset$	$\sqsubset$

Step 4: entailment projection

P	*Jimmy Dean*	*refused to*			*move*	*without*	*blue*	*jeans*
H	*James Dean*		*did*	*n't*	*dance*	*without*		*pants*
edit index	1	2	3	4	5	6	7	8
edit type	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
lex feats	strsim= 0.67	implic: +/o	cat:aux	cat:neg	hypo			hyper
lex entrel	$\equiv$	$\mid$	$\equiv$	$\wedge$	$\sqsupset$	$\equiv$	$\sqsubset$	$\sqsubset$
project-ivity	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$	$\uparrow$
atomic entrel	$\equiv$	$\mid$	$\equiv$	$\wedge$	$\sqsubset$	$\equiv$	$\sqsubset$	$\sqsubset$

inversion

Step 5: Entailment composition

P	*Jimmy Dean*	*refused to*			*move*	*without*	*blue*	*jeans*
H	*James Dean*		*did*	*n't*	*dance*	*without*		*pants*
edit index	1	2	3	4	5	6	7	8
edit type	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
lex feats	strsim= 0.67	implic: +/o	cat:aux	cat:neg	hypo			hyper
lex entrel	$\equiv$	$\mid$	$\equiv$	$\wedge$	$\sqsupset$	$\equiv$	$\sqsubset$	$\sqsubset$
project-ivity	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$	$\uparrow$
atomic entrel	$\equiv$	$\mid$	$\equiv$	$\wedge$	$\sqsubset$	$\equiv$	$\sqsubset$	$\sqsubset$
compo-sition	$\equiv$	$\mid$	$\mid$	$\sqsubset$	$\sqsubset$	$\sqsubset$	$\sqsubset$	$\sqsubset$

interesting

final answer

Outline

- Introduction
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

The FraCaS test suite

- FraCaS: mid-90s project in computational semantics
- 346 “textbook” examples of textual inference 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

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

gold		*yes*	*no*	*unk*	total
	yes	67	4	31	102
	no	1	16	4	21
	unk	7	7	46	60
	total	75	27	81	183

Outline

- Introduction
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

The RTE3 test suite

- RTE: more “natural” textual 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	*Dogs*	*hate*	*figs*
Dogs	1.00	0.00	0.33
do	0.67	0.00	0.00
n't	0.33	0.25	0.00
like	0.00	0.25	0.25
fruit	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 sim)^IDF

$= \Pi_h$ P(h|P)

A simple bag-of-words model

P \ H	*Dogs*	*hate*	*figs*	max	IDF	P(p|H)	P(P|H)
Dogs	1.00	0.00	0.33	1.00	0.43	1.00	
do	0.67	0.00	0.00	0.67	0.11	0.96	
n't	0.33	0.25	0.00	0.33	0.05	0.95	0.43
like	0.00	0.25	0.25	0.25	0.25	0.71	
fruit	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		$= \Pi_h$ 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 $\sqsupset$ 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
	test	56.1	63.0	69.0	63.4
BoW + NatLog + other	dev	52.7	70.9	72.6	70.5
	test	58.7	63.0	72.2	64.0

+13 probs

+8 probs

Outline

- Introduction
- Background on natural logic & monotonicity
- A new(ish) model of natural logic
- The NatLog system
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

What natural logic can't do

- Not a universal solution for textual inference
- Many types of inference not amenable to natural logic
  - Paraphrase: *Eve was let go* ≡ *Eve lost her job*
  - Verb/frame alternation: *he drained the oil* ⊏ *the oil drained*
  - Relation extraction: *Aho, a trader at UBS…* ⊏ *Aho works for UBS*
  - Common-sense reasoning: *the sink overflowed* ⊏ *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* ≡ *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 FOL.
- The NatLog system successfully handles a broad range of such inferences, as demonstrated on the FraCaS test suite.
- Ultimately, open-domain textual inference is likely to require combining disparate reasoners, and a facility for natural logic is a good candidate to be a component of such a system.

:-) Thanks! Questions?