Textual Inference

Bill MacCartney
CS224U: Natural Language Understanding
Stanford University
16 February 2012

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 soon…*
  - Relation extraction (database building)
  - Document summarization

A two-part talk

1. The Stanford RTE system
   - Describes a system to which I was one of many contributors
   - Starts by aligning dependency trees of premise & hypothesis
   - Then extracts global, semantic features and classifies entailment
   - Based on talk I presented at NAACL-06 (with updated results)

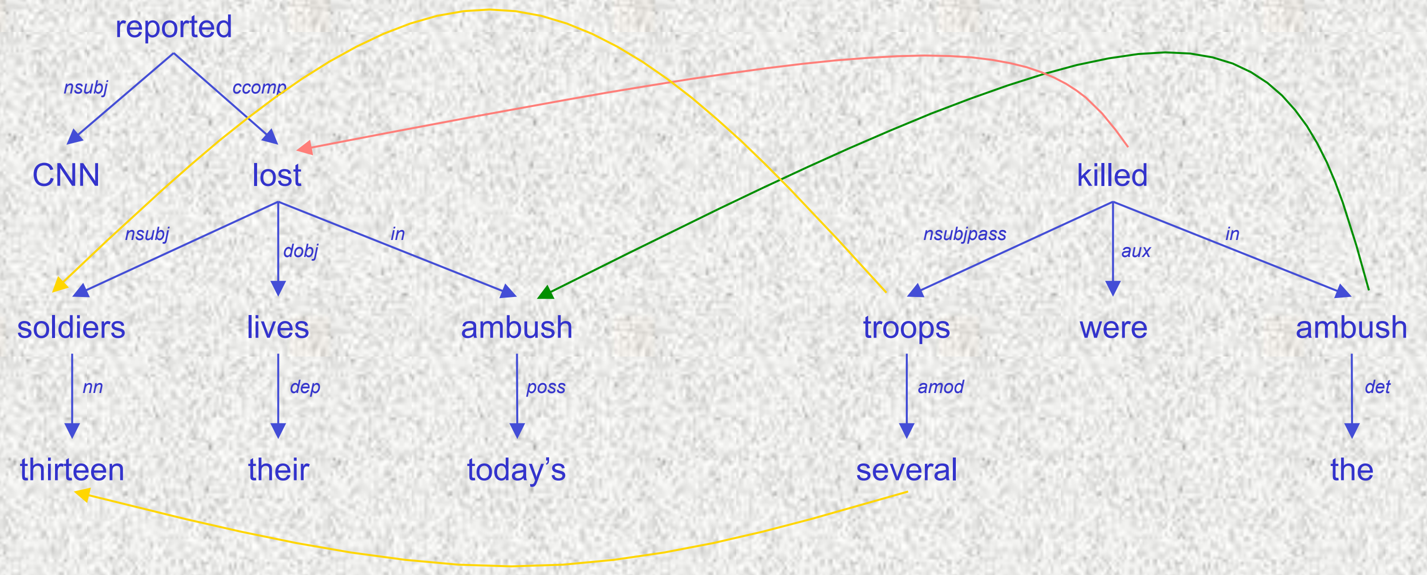
2. The NatLog system: natural logic for textual inference
   - Describes a system which I developed in my dissertation work
   - Assumes an alignment, but interprets as an edit sequence
   - Classifies entailments across each edit & composes results
   - Based on a talk I presented at COLING-08

Textual inference as graph alignment

- Many efforts have converged on this approach
  [Haghighi et al. 05, de Salvo Braz et al. 05]
- Represent *P* & *H* as typed dependency graphs
  - Graph nodes = words of sentence
  - Graph edges = grammatical relations (subject, possessive, etc.)
- Find least-cost alignment of *H* to (part of) *P*
  - Can *H* be (approximately) embedded within *P*?
- Use locally-decomposable cost model
  - Lexical costs penalize aligning semantically unrelated words
  - Structural costs penalize aligning dissimilar subgraphs
- Assume good alignment ⇒ valid inference

Example: graph alignment

P: *CNN reported that thirteen soldiers lost their lives in today's ambush.* ⊨ H: *Several troops were killed in the ambush.*



Problems with alignment models

- Alignments are important, but…
- Good alignment ⇎ valid inference:
  1. Assumption of upward monotonicity
  2. Assumption of locality
  3. Confounding of alignment and entailment

Problem 1: non-monotonicity

- In normal "upward monotone" contexts, broadening a concept preserves truth:

  P: *Some Korean historians believe the murals are of Korean origin.* ⊨
  H: *Some historians believe the murals are of Korean origin.*

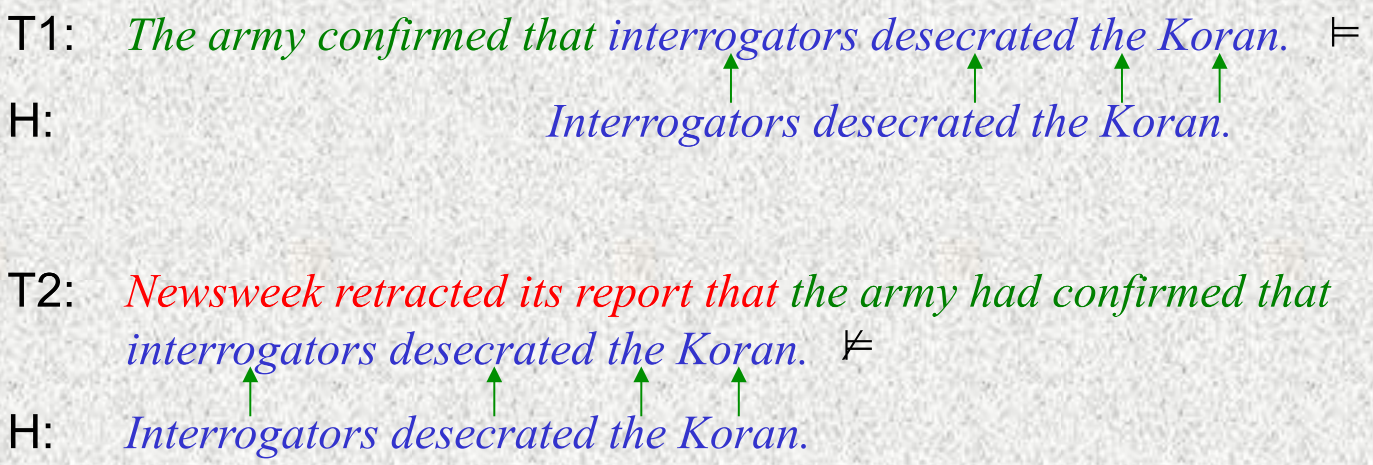
- But not in "downward monotone" contexts:

  P: *Few Korean historians doubt that Koguryo belonged to Korea.* ⊭
  H: *Few historians doubt that Koguryo belonged to Korea.*

- Lots of constructs invert monotonicity!
  - explicit negation: *not*
  - restrictive quantifiers: *no*, *few*, *at most* n
  - negative or restrictive verbs: *lack*, *fail*, *deny*
  - preps & adverbs: *without*, *except*, *only*
  - comparatives and superlatives
  - antecedent of a conditional: *if*

Problem 2: non-locality

- To be tractable, alignment scoring must be local
- But valid inference can hinge on non-local factors:



T1: *The army confirmed that interrogators desecrated the Koran.* ⊨

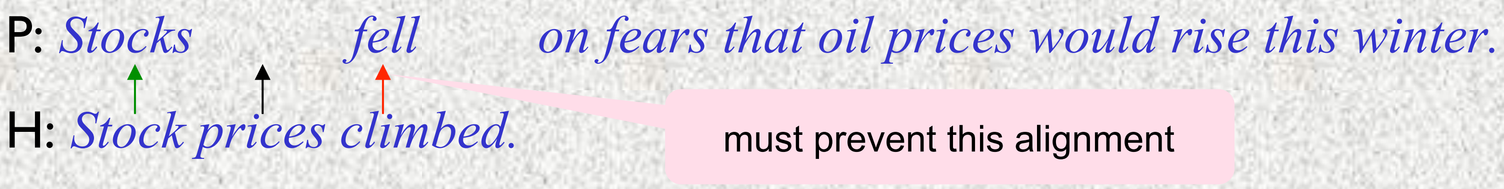
H: *Interrogators desecrated the Koran.*

T2: *Newsweek retracted its report that the army had confirmed that interrogators desecrated the Koran.* ⊭

H: *Interrogators desecrated the Koran.*

Problem 3: confounding alignment & inference

- If alignment ⇒ entailment, lexical cost model *must* penalize e.g. antonyms, inverses:

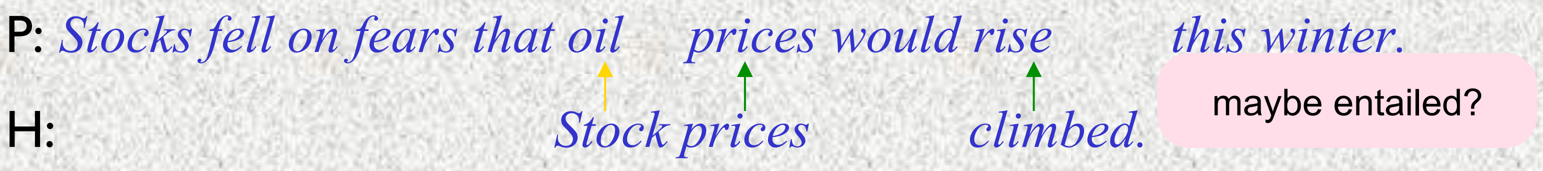


P: *Stocks fell on fears that oil prices would rise this winter.*

H: *Stock prices climbed.*

must prevent this alignment

- But aligner will seek the *best* alignment:



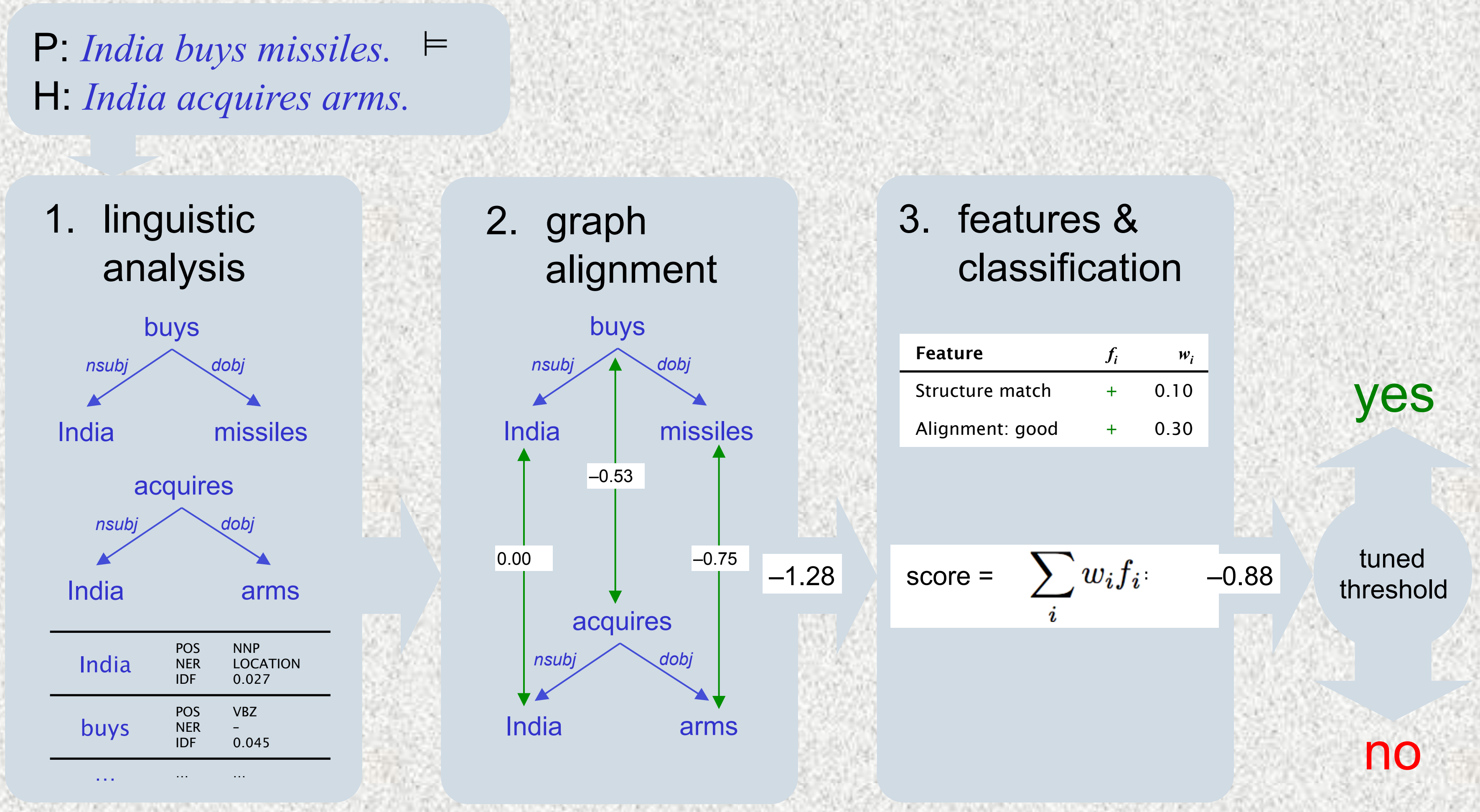
P: *Stocks fell on fears that oil prices would rise this winter.*

H: *Stock prices climbed.*

maybe entailed?

- Actually, we want the first alignment, and then a separate assessment of entailment! [cf. Marsi & Krahmer 05]

Solution: three-stage architecture



P: *India buys missiles.* ⊨
H: *India acquires arms.*

1. linguistic analysis

India	POS NER IDF	NNP LOCATION 0.027
buys	POS NER IDF	VBZ – 0.045
…	…	…

2. graph alignment

3. features & classification

Feature	f_i	w_i
Structure match	+	0.10
Alignment: good	+	0.30

score = $\sum_i w_i f_i$: –0.88

1. Linguistic analysis

- Typed dependencies from statistical parser [de Marneffe et al. 06]
- Collocations from WordNet *(Bill hung_up the phone)*
- Statistical named entity recognizers [Finkel et al. 05]
- Canonicalization of quantity, date, and money expressions

  P: *Kessler's team conducted 60,643* [60,643] *face-to-face interviews…* ⊨

  H: *Kessler's team interviewed more than 60,000* [>60,000] *adults…*
- Semantic role identification: PropBank roles [Toutanova et al. 05]
- Coreference resolution:

  P: *Since its formation in 1948, Israel…* ⊨ H: *Israel was established in 1948.*
- Hand-built: acronyms, country and nationality, factive verbs
- TF-IDF scores

2. Aligning dependency graphs

- Beam search for least-cost alignment
- Locally decomposable cost model
  - Can't do Viterbi-style DP or heuristic search without this
  - Assessment of global features postponed to next stage
- Lexical matching costs
  - Use lexical semantic relatedness scores derived from WordNet, LSA, string sim, distributional similarity [Lin 98]
  - Do *not* penalize antonyms, inverses, alternatives…
- Structural matching costs
  - Each edge in graph of *H* is projected to a path in graph of *P*
  - Preserved edges get low cost; longer paths cost more

3. Features of valid inferences

- After alignment, extract features of inference
  - Look for *global* characteristics of valid and invalid inferences
  - Features embody crude semantic theories
  - Feature categories: adjuncts, modals, quantifiers, implicatives, antonymy, tenses, pred-arg structure, explicit numbers & dates
  - Alignment score is also an important feature
- Extracted features ⇒ statistical model ⇒ score
  - Can learn feature weights using logistic regression
  - Or, can use hand-tuned weights
- (Score ≥ threshold) ? ⇒ prediction: yes/no
  - Threshold can be tuned

Features: restrictive adjuncts

- Does hypothesis add/drop a restrictive adjunct?
  - Adjunct is dropped: usually truth-preserving
  - Adjunct is added: suggests no entailment
  - But in a *downward monotone* context, this is reversed

P: *In all, Zerich bought $422 million worth of oil from Iraq, according to the Volcker committee.* $\not\models$

H: *Zerich bought oil from Iraq during the embargo.*

P: *Zerich didn't buy any oil from Iraq, according to the Volcker committee.* $\models$

H: *Zerich didn't buy oil from Iraq during the embargo.*

- Generate features for add/drop, monotonicity

Features: modality

P: *Sharon warns Arafat could be targeted for assassination.* ⊭
H: *Prime minister targeted for assassination.* [RTE1-98]

P: *After the trial, Family Court found the defendant guilty of violating the order.* ⊭
H: *Family Court cannot punish the guilty.* [RTE1-515]

- Define 6 canonical modalities
- Identify modalities of *P* & *H*:

modality	markers
ACTUAL	(default)
NOT_ACTUAL	*not*, *no*, …
POSSIBLE	*could*, *might*, *possibly*, …
NOT_POSSIBLE	*impossible*, *couldn't*, …
NECESSARY	*must*, *has to*, …
NOT_NECESSARY	*might not*, …

- Map ⟨*P*, *H*⟩ modality pairs to categorical features:

premise	hypothesis	feature
ACTUAL	POSSIBLE	good
NECESSARY	NOT_ACTUAL	bad
POSSIBLE	ACTUAL	neutral
…	…	…

Features: factives & implicatives

P: *Libya has tried, with limited success, to develop its own indigenous missile, and to extend the range of its aging SCUD force for many years under the Al Fatah and other missile programs.* ⊭

H: *Libya has developed its own domestic missile program.*

- Evaluate governing verbs for implicativity class
  - Unknown: *say*, *tell*, *suspect*, *try*, *…*
  - Fact: *know*, *acknowledge*, *ignore*, *…*
  - True: *manage to*, *…*
  - False: *fail to*, *forget to*, *…*

[cf. Nairn et al. 06]

- Need to check for ↓-monotone context here too
  - *not try to win* ⊭ *not win*, but *not manage to win* ⊨ *not win*

Evaluation: PASCAL RTE

- RTE = recognizing textual "entailment" [Dagan et al. 05]
- Does premise *P* "entail" hypothesis *H*?

  P: *Wal-Mart defended itself in court today against claims that its female employees were kept out of jobs in management because they are women.* ⊨

  H: *Wal-Mart was sued for sexual discrimination.*
- Three annual competitions (so far)
  - RTE1 (2005): 567 dev pairs, 800 test pairs
  - RTE2 (2006) and RTE3 (2007): 800 dev pairs, 800 test pairs
- Considerable variance from year to year
- High inter-annotator agreement (~95%)

Results & useful features

RTE1 test set (800 pairs)

Algorithm	*Acc.*	*CWS**
Random	50.0	50.0
Jijkoun & de Rijke 05	55.3	55.9
Bos & Markert 05 (strict)	57.7	63.2
Alignment only	54.5	59.7
Learned weights	**59.1**	**63.9**
Hand-tuned weights	**59.1**	65.0

*confidence-weighted score (standard RTE1 evaluation metric)

Most useful features

Positive

- Added adjunct in ↓ context
- Pred-arg structure match
- Modal: yes
- premise is embedded in factive
- Good alignment score

Negative

- Date inserted/mismatched
- Pred-arg structure mismatch
- Quantifier mismatch
- Bad alignment score
- Different polarity
- Modal: no/don't know

Results for all RTE data [updated]

RTE1	dev	test
bag of words	54.0	53.6
Stanford (hand-tuned)	60.3	59.1
Stanford (learned)	61.2	59.1

RTE2	dev	test
bag of words	57.0	57.6
Stanford (hand-tuned)	67.0	58.3
Stanford (learned)	66.9	60.5

RTE3	dev	test
bag of words	68.9	63.0
Stanford (core)	67.3	60.5
Stanford (+NatLog)	69.6	63.6

+25 probs

What we have trouble with

- Non-entailment is easier than entailment
  - Good at finding knock-out features
  - But, hard to be certain that we've considered everything
- Lots of adjuncts, but which are restrictive?

  H: *Maurice was subsequently killed in Angola.*
- Multiword "lexical" semantics/world knowledge
  - We're pretty good at synonyms, hyponyms, antonyms
  - But we aren't good at recognizing multi-word equivalences

  P: *David McCool took the money and decided to start Muzzy Lane in 2002.*

  H: *David McCool is the founder of Muzzy Lane.* [RTE2-379]
  - Other teams (e.g. LCC) have done well with paraphrase models

Conclusion [of the first part!]

- Alignment models promising, but flawed:
  1. Assumption of monotonicity
  2. Assumption of locality
  3. Confounding of alignment and inference
- Solution: align, *then* judge validity of inference
- We extract global-level semantic features
  - Working from richly-annotated, aligned dependency graphs … not just word sequences
  - Features are designed to embody crude semantic theories
- Still lots of room to improve…

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

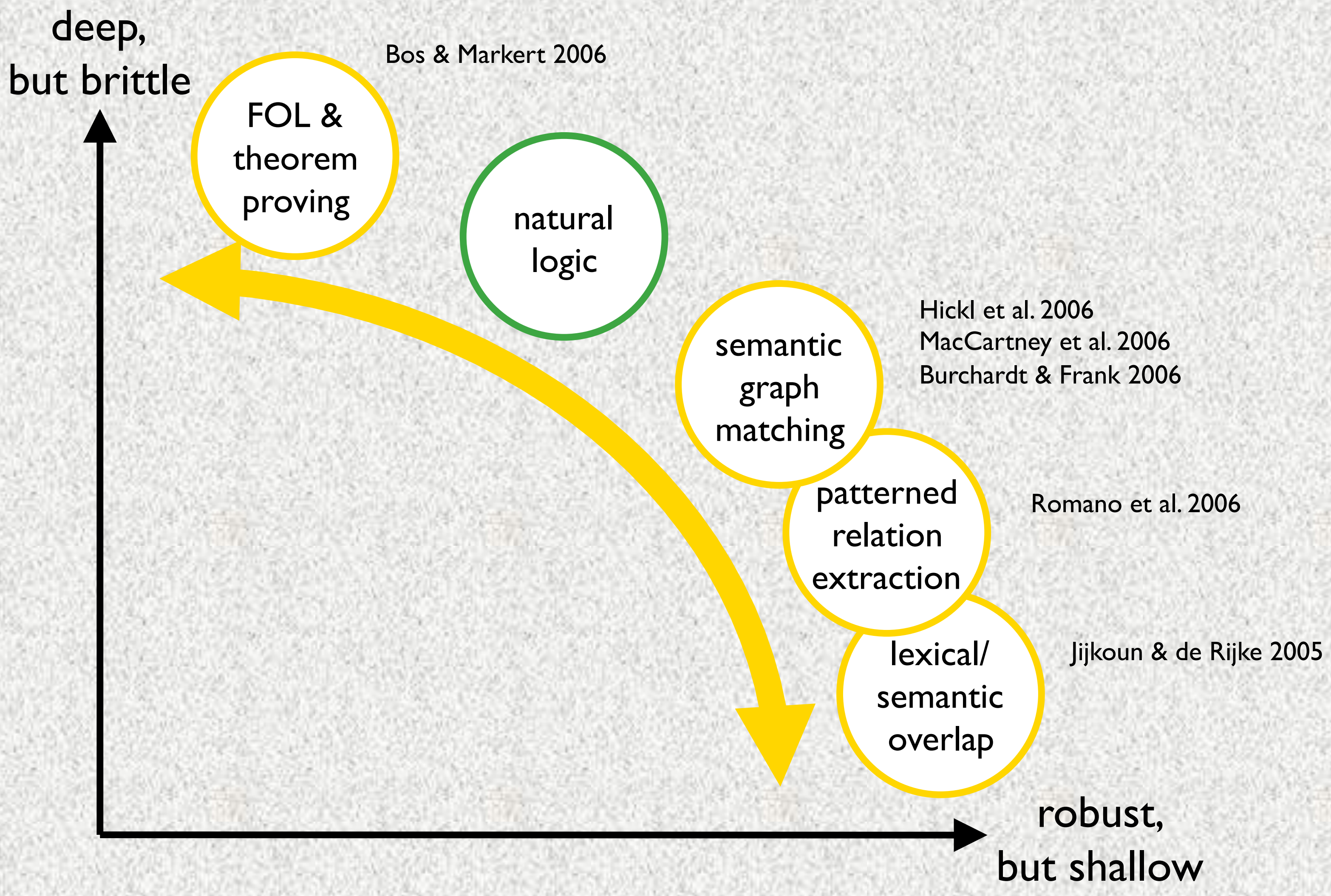
No state completely forbids casino gambling for kids.

No state restricts gambling.

No state or city completely forbids casino gambling.

What kind of textual inference system could predict this?

Textual inference: a spectrum of approaches



deep, but brittle

FOL & theorem proving

Bos & Markert 2006

natural logic

semantic graph matching

Hickl et al. 2006
MacCartney et al. 2006
Burchardt & Frank 2006

patterned relation extraction

Romano et al. 2006

lexical/ semantic overlap

Jijkoun & de Rijke 2005

robust, but shallow

What is natural logic?

- (natural logic ≠ 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: → ¬ ∧ ∨ ∀ ∃
- 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, …

Monotonicity calculus (Sánchez Valencia 1991)

- Entailment as semantic containment:

  rat < *rodent*, *eat* < *consume*, *this morning* < *today*, *most* < *some*

- Monotonicity classes for semantic functions
  - Upward monotone: *some rats dream* < *some rodents dream*
  - Downward monotone: *no rats dream* > *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, …)

  Garfield is a cat < *Garfield is not a dog*

Implicatives & factives

- Work at PARC, esp. Nairn et al. 2006
- Explains inversions & nestings of implicatives & factives
  - *Ed did not forget to force Dave to leave* ⇒ *Dave left*
- Defines 9 implication signatures
- "Implication projection algorithm"
  - Bears some resemblance to monotonicity calculus
- But, fails to connect to containment or monotonicity
  - *John refused to dance* ⇒ *John didn't tango*

Outline

- Introduction
- Foundations of Natural Logic
- The NatLog System
- Experiments with FraCaS
- Experiments with RTE
- Conclusion

A new theory of natural logic

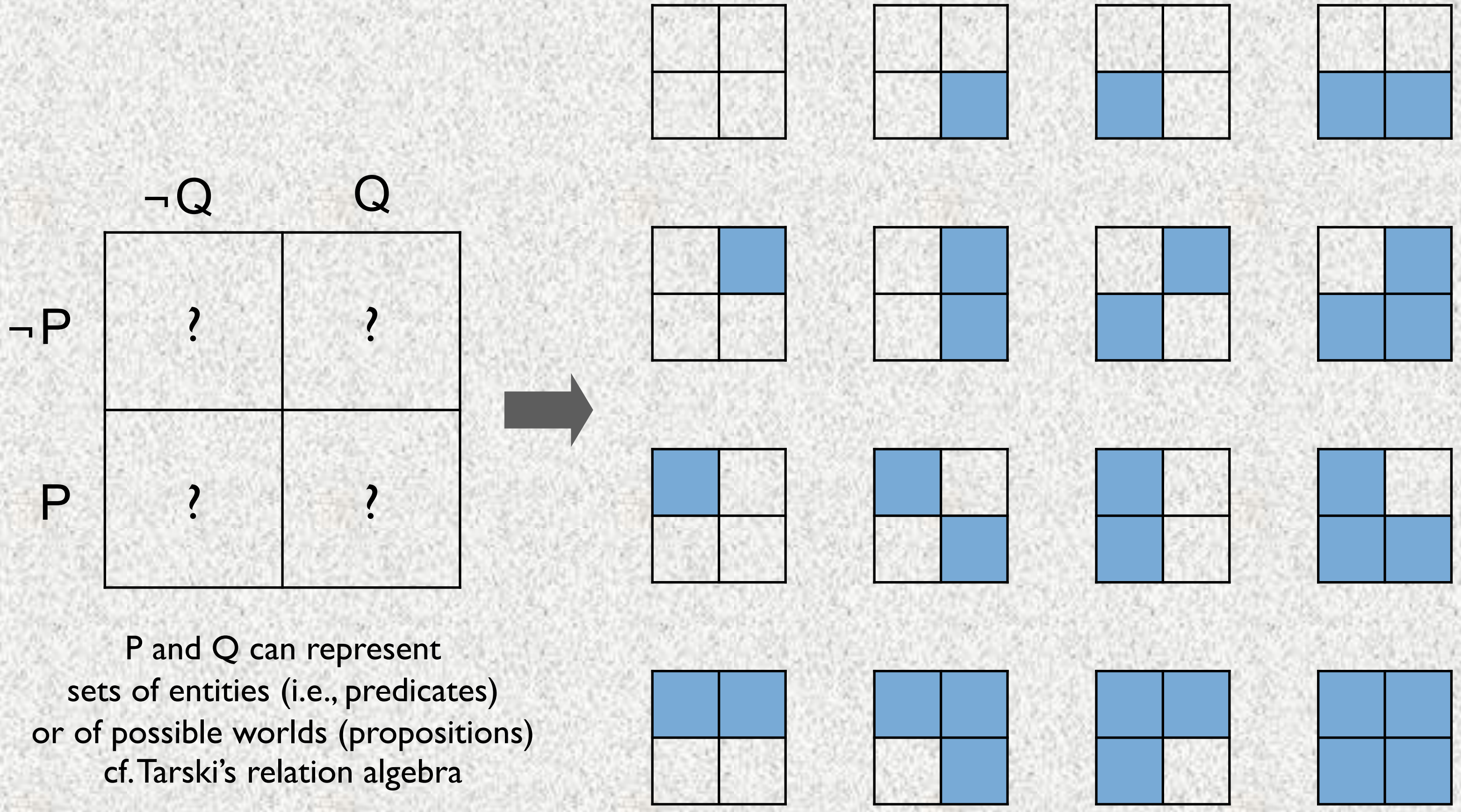
Three elements:

1. an inventory of entailment relations
   - semantic containment relations of Sánchez Valencia
   - plus semantic exclusion relations
2. a concept of projectivity
   - explains entailments compositionally
   - generalizes Sánchez Valencia's monotonicity classes
   - generalizes Nairn et al.'s implication signatures
3. a weak proof procedure
   - composes entailment relations across chains of edits

Entailment relations in past work

<table>
<tr><th></th><th>$\frac{X \text{ is a couch}}{X \text{ is a sofa}}$</th><th>$\frac{X \text{ is a crow}}{X \text{ is a bird}}$</th><th>$\frac{X \text{ is a fish}}{X \text{ is a carp}}$</th><th>$\frac{X \text{ is a hippo}}{X \text{ is hungry}}$</th><th>$\frac{X \text{ is a man}}{X \text{ is a woman}}$</th></tr>
<tr><td>2-way
RTE1,2,3</td><td colspan="2">Yes
entailment</td><td colspan="3">No
non-entailment</td></tr>
<tr><td>3-way
RTE4, FraCaS, PARC</td><td colspan="2">Yes
entailment</td><td colspan="2">Unknown
non-entailment</td><td>No
contradiction</td></tr>
<tr><td>containment
Sánchez-Valencia</td><td>P = Q
equivalence</td><td>P < Q
forward entailment</td><td>P > Q
reverse entailment</td><td colspan="2">P # Q
non-entailment</td></tr>
</table>

16 elementary set relations

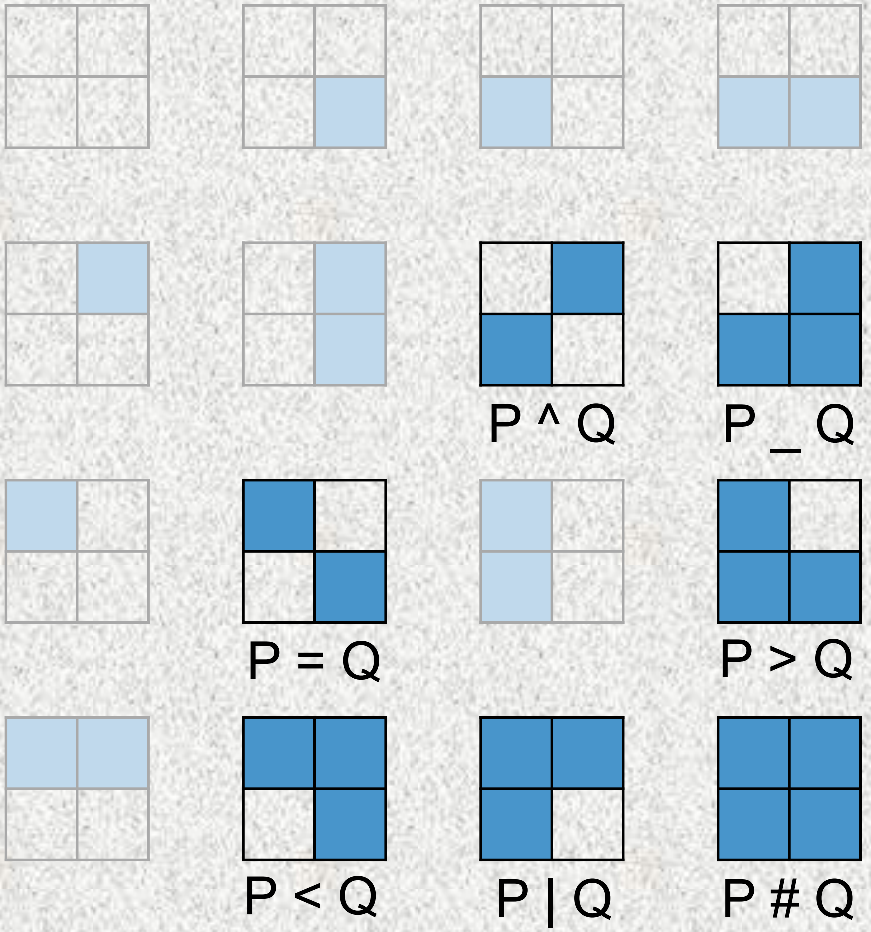


	¬Q	Q
¬P	?	?
P	?	?

P and Q can represent
sets of entities (i.e., predicates)
or of possible worlds (propositions)
cf. Tarski's relation algebra

16 elementary set relations

	¬Q	Q
¬P	?	?
P	?	?



P and Q can represent
sets of entities (i.e., predicates)
or of possible worlds (propositions)
cf. Tarski's relation algebra

7 basic entailment relations

symbol	name	example	2-way	3-way
P = Q	equivalence	*couch* = *sofa*	yes	yes
P < Q	forward (strict)	*crow* < *bird*	yes	yes
P > Q	reverse (strict)	*European* > *French*	no	unk
P ^ Q	negation (exhaustive exclusion)	*human* ^ *nonhuman*	no	no
P | Q	alternation (non-exhaustive exclusion)	*cat* | *dog*	no	no
P _ Q	cover (non-exclusive exhaustion)	*animal* _ *nonhuman*	no	unk
P # Q	independence	*hungry* # *hippo*	no	unk

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*

Projectivity (= monotonicity++)

- How do the entailments of a compound expression depend on the entailments of its parts?
- How does the entailment relation between (*f x*) and (*f y*) depend on the entailment relation between *x* and *y* (and the properties of *f*)?
- Monotonicity gives partial answer (for =, <, >, #)
- But what about the other relations (^, |, _)?
- We'll categorize semantic functions based on how they *project* the basic entailment relations

Example: projectivity of *not*

projection			example
=	→	=	*not happy* = *not glad*
<	→	>	*didn't kiss* > *didn't touch*
>	→	<	*isn't European* < *isn't French*
#	→	#	*isn't swimming* # *isn't hungry*
^	→	^	*not human* ^ *not nonhuman*
|	→	_	*not French* _ *not German*
_	→	|	*not more than 4* | *not less than 6*

downward monotonicity

swaps these too

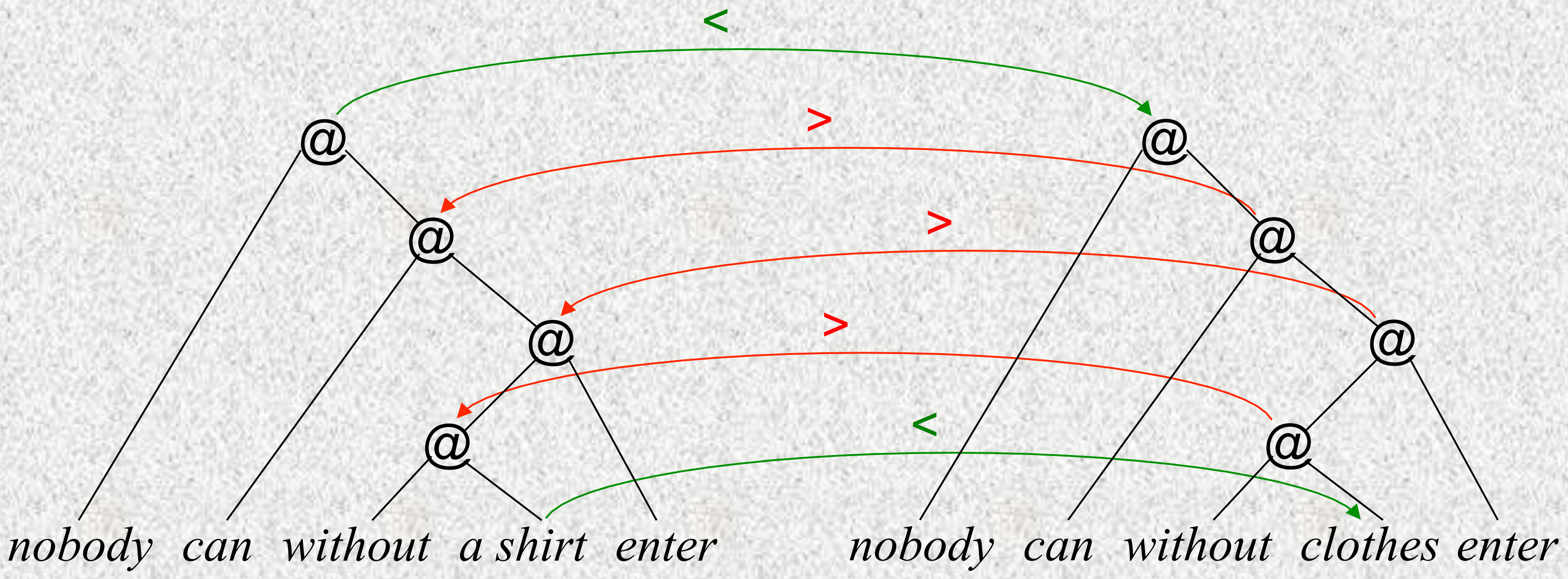
Example: projectivity of *refuse*

	projection	example
	= → =	
	< → >	*refuse to tango* > *refuse to dance*
downward monotonicity	> → <	
	# → #	
switch	^ → |	*refuse to stay* | *refuse to go*
	| → #	*refuse to tango* # *refuse to waltz*
blocks, not swaps	_ → #	

Projecting entailment relations upward

Nobody can enter without a shirt < *Nobody can enter without clothes*

- Assume idealized semantic composition trees
- Propagate lexical entailment 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, …
2. Determine lexical entailment 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 entailment relation across each edit
4. Compose entailment relations across sequence of edits

Composing entailment relations

- Relation composition: if *a R b* and *b S c*, then *a ? c*
  - cf. Tarski's relation algebra
- Many compositions are intuitive

  = ° = ⇒ =   < ° < ⇒ <   < ° = ⇒ <   ^ ° ^ ⇒ =

- Some less obvious, but still accessible

  | ° ^ ⇒ <   *fish* | *human*, *human* ^ *nonhuman*, *fish* < *nonhuman*

- But some yield *unions* of basic entailment relations!

  | ° | ⇒ ⋃{=, <, >, |, #}   (i.e. the non-exhaustive relations)

  - Larger unions convey less information (can approx. with #)
  - This limits power of proof procedure described

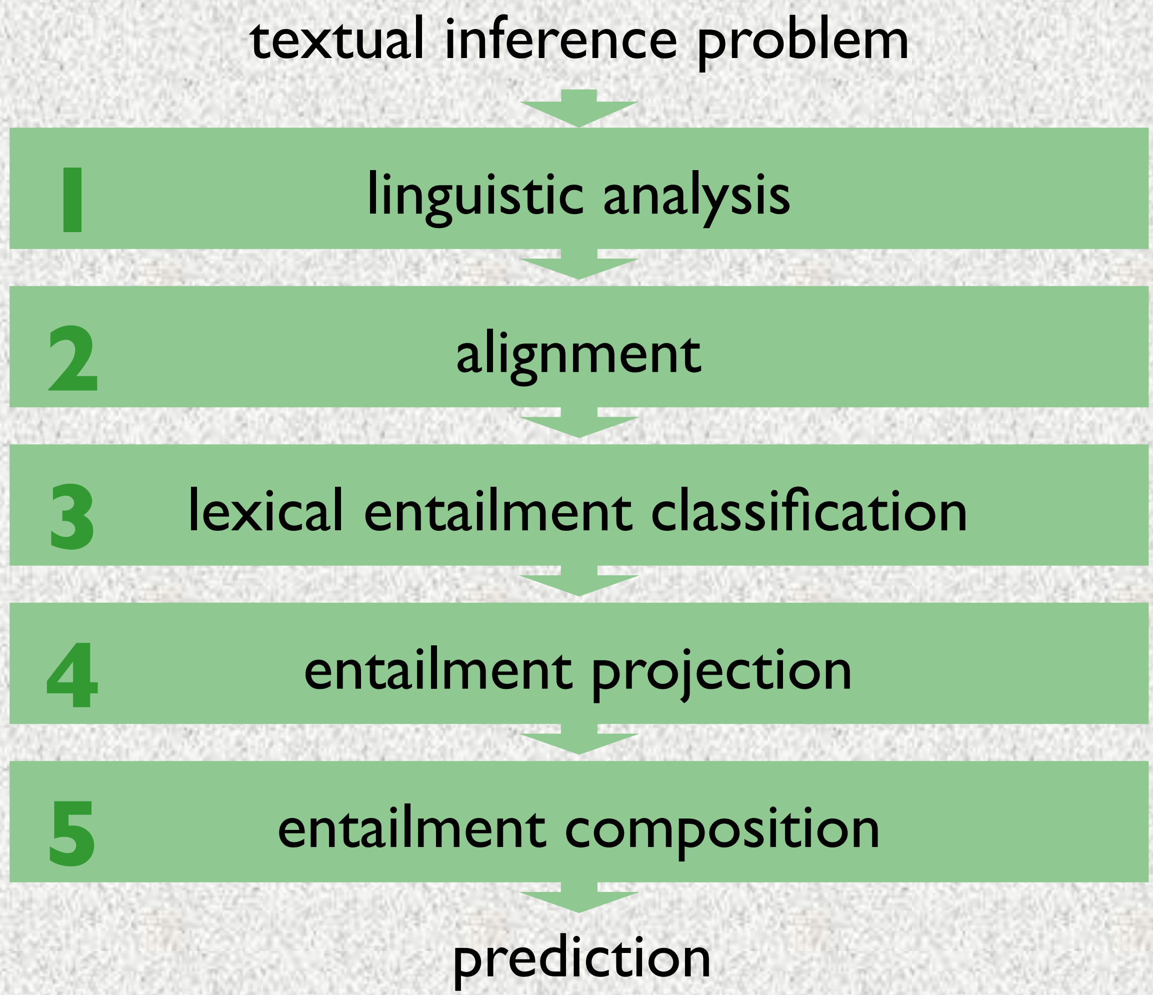
Implicatives & factives

- Nairn et al. 2006 define nine *implication signatures*
- These encode implications (+, –, o) in + and – contexts
  - *Refuse* has signature –/o:
    refuse to dance implies *didn't dance*
    didn't refuse to dance implies neither *danced* nor *didn't dance*
- Signatures generate different relations when deleted
  - Deleting –/o generates |
    Jim refused to dance | *Jim danced*
    Jim didn't refuse to dance _ *Jim didn't dance*
  - Deleting o/– generates <
    Jim attempted to dance < *Jim danced*
    Jim didn't attempt to dance > *Jim didn't dance*
- (Factives are only partly explained by this account)

Outline

- Introduction
- Foundations 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

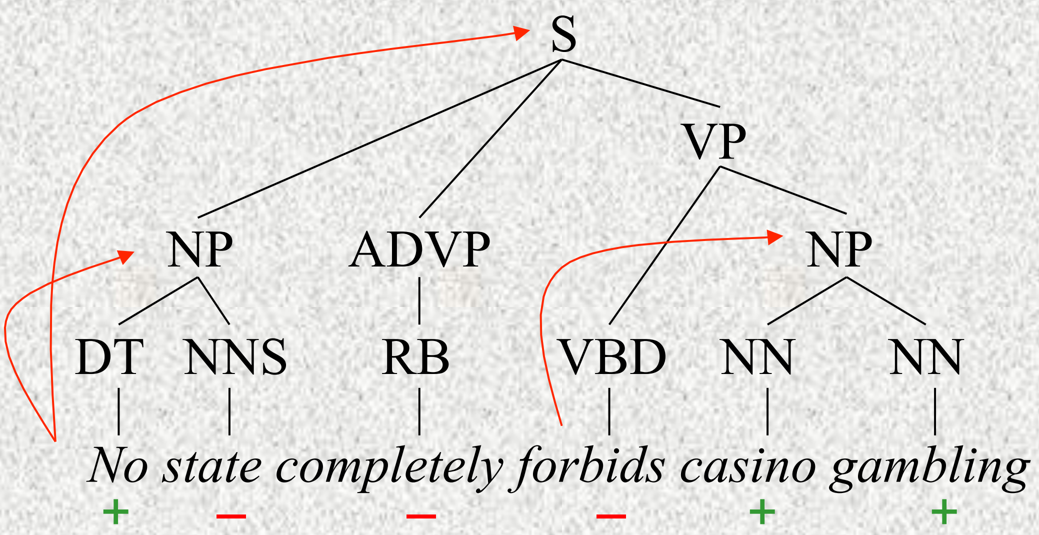
4 entailment projection

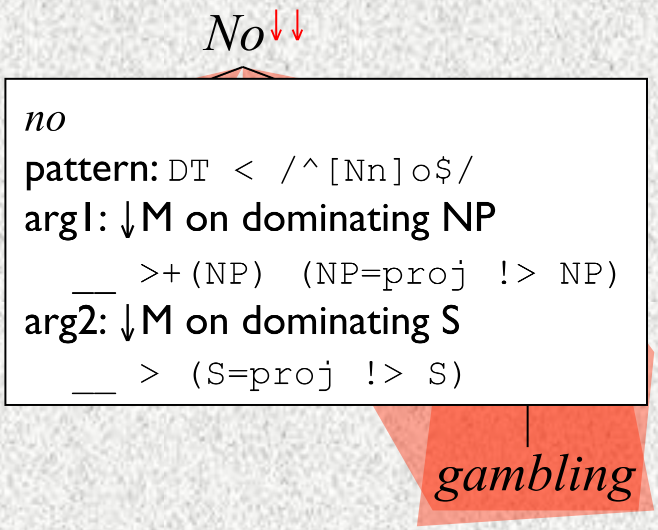
5 entailment composition

prediction

Step 1: Linguistic analysis

- Tokenize & parse input sentences (future: & NER & coref & …)
- Identify items w/ special projectivity & determine scope
- Problem: PTB-style parse tree ≠ semantic structure!





no

pattern: `DT < /^[Nn]o$/`

arg1: ↓M on dominating NP

`__ >+(NP) (NP=proj !> NP)`

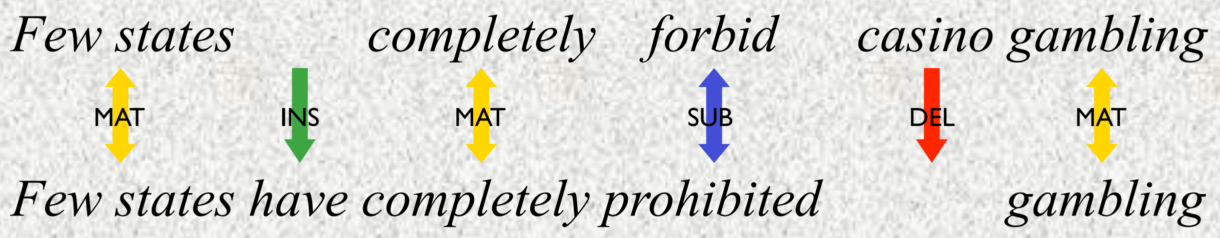
arg2: ↓M on dominating S

`__ > (S=proj !> S)`

- 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



Few states completely forbid casino gambling

Few states have completely prohibited gambling

- 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 entailment 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
  - >99% accuracy on training data — 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	=	|	=	^	>	=	<	<

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	=	|	=	^	>	=	<	<
project-ivity	↑	↑	↑	↑	↓	↓	↑	↑
atomic entrel	=	|	=	^	<	=	<	<

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	=	|	=	^	>	=	<	<
project-ivity	↑	↑	↑	↑	↓	↓	↑	↑
atomic entrel	=	|	=	^	<	=	<	<
compo-sition	=	|	|	<	<	<	<	<

interesting

final answer

Outline

- Introduction
- Foundations 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
current system	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
this work	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			
		yes	*no*	*unk*	total
gold	*yes*	67	4	31	102
	no	1	16	4	21
	unk	7	7	46	60
	total	75	27	81	183

Outline

- Introduction
- Foundations 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 ⇒ propagation of errors from atomic model

RTE3 examples

P *As leaders gather in Argentina ahead of this weekends 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	similarity scores on [0, 1] for each pair of words (I used a really simple-minded similarity function based on Levenshtein string-edit distance)
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	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		= Π_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	0.43
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	
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		= Π_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 & NatLog

- MaxEnt classifier
- BoW features: P(H|P), P(P|H)
- NatLog features:
  7 boolean features encoding predicted entailment 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 < or = for all but one or two edits
  - But NatLog predicts some other relation for remaining edits!
  - Most commonly, it predicts > for an insertion (e.g., RTE3_dev.71)
  - 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 < or =

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

(each data set contains 800 problems)

Outline

- Introduction
- Foundations 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?