Natural Logic for Textual Inference

MacCartney and Manning

Gareth Dwyer

Outline

- Introduction
  - Shallow/Robust - Deep/Brittle tradeoff
  - Limits of Natural Logic
- Natural Logic Foundations
  - Entailment and Monotonicity
  - Composition
- NatLog System
  - Preprocessing, Alignment, Entailment classification
- Fracas Experiments
  - FraCaS test suite
  - Results
- RTE Experiments
  - RTE Test Suite
  - Results
- Conclusion

Theory and Practice

- Theory vs Practice?

Theory and Practice

- Theory vs Practice?
- In theory, nothing works - but everyone knows why

Theory and Practice

- Theory vs Practice?
- In theory, nothing works - but everyone knows why
- In practice, everything works, but no-one knows why

Theory and Practice

- Theory vs Practice?
- In theory, nothing works - but everyone knows why
- In practice, everything works, but no-one knows why
- Combination of theory and practice?

Theory and Practice

- Theory vs Practice?
- In theory, nothing works - but everyone knows why
- In practice, everything works, but no-one knows why
- Combination of theory and practice?
  - Nothing works and no-one knows why...

Deep Logic or Shallow Semantics

- Textual Inference
  - Can a hypothesis be inferred from a premise?
- Existing systems
  - First Order Logic
    - Complex proofs possible
    - Precise, but ‘brittle’
    - Need to translate to FOL. High precision, low recall.
  - Shallow semantics
    - Lexical/Semantic overlap
    - Robust, but imprecise.
    - “No case of indigenously acquired rabies infection has been confirmed in the past 2 years” -/-> “No rabies cases have been confirmed”

Middle ground?

- Natural Language < **Natural Logic** < First Order Logic
  - No Logical Notation or Model Theory
  - Proofs by “Incremental Edits”
  - Inference Rules for Semantic contractions and expansions
    - Truth preservation

- “**Precise** reasoning about *monotonicity*, while sidestepping the difficulties of translating sentences to FOL”
  - Logical precision without logic?

What doesn't Natural Logic do?

- Kinds of inference not covered by Natural Logic
  - Temporal reasoning
  - Causal reasoning
    - Sold nuclear plans -> Possessed nuclear plans
  - Paraphrasing
    - Flew to Rome -> Took a flight to Rome
  - Relation Extraction
    - Bill Gates and his wife Melinda -> Melinda Gates is married to Bill Gates
  - Deep proof searching
- But can integrate with other systems

Ax. 1. $\{P(\varphi) \wedge \Box\, \forall x[\varphi(x) \rightarrow \psi(x)]\} \rightarrow P(\psi)$
Ax. 2. $P(\neg\varphi) \leftrightarrow \neg P(\varphi)$
Th. 1. $P(\varphi) \rightarrow \Diamond\, \exists x[\varphi(x)]$
Df. 1. $G(x) \iff \forall\varphi[P(\varphi) \rightarrow \varphi(x)]$
Ax. 3. $P(G)$
Th. 2. $\Diamond\, \exists x\, G(x)$
Df. 2. $\varphi \text{ ess } x \iff \varphi(x) \wedge \forall\psi\{\psi(x) \rightarrow \Box\, \forall y[\varphi(y) \rightarrow \psi(y)]\}$
Ax. 4. $P(\varphi) \rightarrow \Box\, P(\varphi)$
Th. 3. $G(x) \rightarrow G \text{ ess } x$
Df. 3. $E(x) \iff \forall\varphi[\varphi \text{ ess } x \rightarrow \Box\, \exists y\, \varphi(y)]$
Ax. 5. $P(E)$
Th. 4. $\Box\, \exists x\, G(x)$

Outline

- Introduction
  - Shallow/Robust - Deep/Brittle tradeoff
  - Limits of Natural Logic
- Natural Logic Foundations
  - Entailment and Monotonicity
  - Composition
- NatLog System
  - Preprocessing, Alignment, Entailment classification
- Fracas Experiments
  - FraCaS test suite
  - Results
- RTE Experiments
  - RTE Test Suite
  - Results
- Conclusion

Natural Logic

- Objective: Explain inferences using Monotonicity
  - Constraints can be *expanded* or *contracted salva veritate*
- *“Every meal without wine is a terrible crime”*
  - *Every = some (expansion)*
  - *meal = dinner (contraction)*
  - *wine = drink (expansion)*
  - *terrible crime = crime (expansion)*
- *“Some dinners without drink are crimes”*
- *How do we tell which ones we can expand and which we can contract?*
  - *‘Upward’ and ‘Downward’ Monotonicity (later)*

Entailment

- Easy and useful
  - Penguin $\sqsubseteq$ bird; tiny $\sqsubseteq$ small; French $\sqsubseteq$ European
  - This morning $\sqsubseteq$ today
  - And $\sqsubseteq$ or
  - Everyone $\sqsubseteq$ someone; all $\sqsubseteq$ most $\sqsubseteq$ some
  - Everyone $\sqsubseteq$ Einstein $\sqsubseteq$ some physicist
- But
  - Eat quickly $\sqsubseteq$ Eat
  - Fake Vaccine $\not\sqsubseteq$ Vaccine

Monotonicity

- Upward Monotonicity
  - tango in Paris ⊑ dance in France (tango ⊑ dance, Paris ⊑ France)
  - f(x) ⊑ f(y) for every x ⊑ y
- Downward Monotonicity
  - didn't dance ⊑ didn't tango
  - f(y) ⊑ f(x) for every x ⊑ y
  - not, few, without, lack, fail, prohibit
- No Monotonicity
  - Prettiest Butterfly ⋢ Prettiest Animal & Prettiest Animal ⋢ Prettiest Butterfly
  - Prettiest Butterfly # Prettiest Animal
- Mixed Monotonicity
  - Every fish swims ⊑ Every shark swims; Every shark swims ⊑ every shark moves
  - [Every has downward monotonicity for first arg, upward for second]

NatLog

- Compositionality
  - h = f ∘ g
  - if either *f()* **or** g() is non-monotome then h() is also non-monotome()
  - if the monotonicity of f() and g() is the same, h() is upward-monotome
  - if the monotonicity of f() and g() is different, h() is downward-monotone
- Example
  - pants ⊑ clothes
  - without(clothes) ⊑ without(pants)
  - not(clothes) ⊑ not(pants)
  - not(without(pants)) ⊑ not(without(clothes))
  - not() and without() are both downward monotone; not() ∘ without() is upward monotone

Outline

- Introduction
  - Shallow/Robust - Deep/Brittle tradeoff
  - Limits of Natural Logic
- Natural Logic Foundations
  - Entailment and Monotonicity
  - Composition
- NatLog System
  - Preprocessing, Alignment, Entailment classification
- Fracas Experiments
  - FraCaS test suite
  - Results
- RTE Experiments
  - RTE Test Suite
  - Results
- Conclusion

NatLog

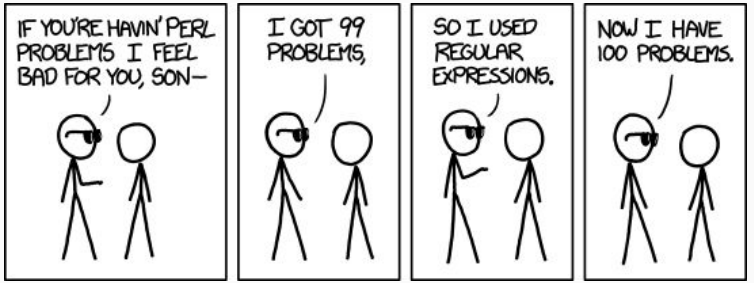
- Three-Stage architecture
  - Linguistic pre-processing
    - Tokenization
    - POS tagging
    - Phrase-Structure
    - **Monotonicity Marking** (Tregex patterns)
  - Textual alignment
    - Atomic edits: Delete; Insert; Substitute; Advance
    - Cost function
  - Entailment classification
    - Sub-entailments from alignment
    - Trained Decision Tree Classifier for atomic entailment classification

Pre-Processing

- Tregex
  - Stanford
  - Matches patterns in trees
- Authors define Tregex patterns manually
  - Downward Monotone
  - Non-Monotone
- Calculate monotonicity of spans
  - Monotonicity composition
  - Marks final monotonicity of span

Alignment

- Match hypothesis to premise through ‘atomic edits’
  - Deletion of span from premise
  - Insertion of span into hypothesis
  - Substitution of premise span into hypothesis
  - Advance (go to next span)
- An Irishman won a Nobel prize” (p) -> “An Irishman won the Nobel prize for literature” (h)
  - An Irishman -> An Irishman (ADV) =
  - won -> won (ADV) =
  - a -> the (SUB) = [light edit]
  - Nobel prize -> Nobel prize (ADV) =
  - -> for literature (INS) ⊐ (reverse)



- Now we have a bunch of smaller entailment problems

Entailment classification

- Now we have a bunch of smaller entailment problems
  - Train classifier to predict relations
  - Features
    - type of edit
    - monotonicity
    - lexical features (inc. WordNet)

relation	symbol	in terms of $\sqsubseteq$	FraCaS	RTE
equivalent	$p = h$	$p \sqsubseteq h, h \sqsubseteq p$	yes	yes
forward	$p \sqsubset h$	$p \sqsubseteq h, h \not\sqsubseteq p$	yes	yes
reverse	$p \sqsupset h$	$h \sqsubseteq p, p \not\sqsubseteq h$	unk	no
independent	$p \# h$	$p \not\sqsubseteq h, h \not\sqsubseteq p$	unk	no
exclusive	$p \mid h$	$p \sqsubseteq \neg h$	no	no

- Composition predictions to get final global entailment prediction
  - $\sqsubset \circ \sqsupset$ is #
  - $= \circ r$ is r
- Nobel prize example again?

Alignment

“An Irishman won a Nobel prize” (p) -> “An Irishman won the Nobel prize for literature” (h)

- An Irishman -> An Irishman (ADV) =
- won -> won (ADV) =
- a -> the (SUB) = [light edit]
- Nobel prize -> Nobel prize (ADV) =
- -> for literature (INS) ⊐ (reverse)

Global entailment = unknown (no)

relation	symbol	in terms of $\sqsubseteq$	FraCaS	RTE
equivalent	$p = h$	$p \sqsubseteq h, h \sqsubseteq p$	yes	yes
forward	$p \sqsubset h$	$p \sqsubseteq h, h \not\sqsubseteq p$	yes	yes
reverse	$p \sqsupset h$	$h \sqsubseteq p, p \not\sqsubseteq h$	unk	no
independent	$p \# h$	$p \not\sqsubseteq h, h \not\sqsubseteq p$	unk	no
exclusive	$p \mid h$	$p \sqsubseteq \neg h$	no	no

Outline

- Introduction
  - Shallow/Robust - Deep/Brittle tradeoff
  - Limits of Natural Logic
- Natural Logic Foundations
  - Entailment and Monotonicity
  - Composition
- NatLog System
  - Preprocessing, Alignment, Entailment classification
- Fracas Experiments
  - FraCaS test suite
  - Results
- RTE Experiments
  - RTE Test Suite
  - Results
- Conclusion

FraCaS experiment

- FraCaS - Manually created test suite
  - 346 problems; 9 categories
  - Questions -> Declarative hypotheses
  - ‘yes’, ‘no’, ‘unk’
- NatLog attempted subset
  - Remove ‘degenerate’ - 12
  - Remove Multiple Ps - 151 (!)
  - Expected poor performance
    - Ellipsis, etc
- Looking for confirmation of adequacy
  - Not that interested in performing well on this specific dataset

fracas-054 answer: unknown

P1	No Scandinavian delegate finished the report on time.
Q	Did any delegate finish the report on time?
H	Some delegate finished the report on time.

fracas-055 answer: yes

P1	Some Irish delegates finished the survey on time.
Q	Did any delegates finish the survey on time?
H	Some delegates finished the survey on time.

FraCaS Results

- Data wasn't 'unseen'
  - But not trained on FraCaS
- Satisfied with performance
  - Applicable sections are good
  - Shows adequacy of model
    - works in theory
- Can't deal with Ellipsis
  - Not expected to
- Accuracy
  - 37/44 Quantifiers is 84.09%

§	Category	Count	% Acc.
1	Quantifiers	44	84.09
2	Plurals	24	41.67
3	Anaphora	6	50.00
4	Ellipsis	25	28.00
5	Adjectives	15	60.00
6	Comparatives	16	68.75
7	Temporal	36	61.11
8	Verbs	8	62.50
9	Attitudes	9	55.56
Applicable sections: 1, 5, 6		75	76.00
All sections		183	59.56

FraCaS Results Discussion

- Confusion Matrix
  - Guess ‘unk’; real answer ‘yes’
    - Bias towards yes in test data
    - System gives ‘unk’ when confused
    - ‘yes’ iff *all* atomics are ⊏ or =
  - Guess ‘yes’ real answer ‘unk’
    - Former student ⊈student
    - Could do better with training examples

	guess			
answer	*yes*	*unk*	*no*	total
yes	62	40	–	102
unk	15	45	–	60
no	6	13	2	21
total	90	91	2	183

Outline

- ❏ Introduction
  - ❏ Shallow/Robust - Deep/Brittle tradeoff
  - ❏ Limits of Natural Logic
- ❏ Natural Logic Foundations
  - ❏ Entailment and Monotonicity
  - ❏ Composition
- ❏ NatLog System
  - ❏ Preprocessing, Alignment, Entailment classification
- ❏ Fracas Experiments
  - ❏ FraCaS test suite
  - ❏ Results
- ❏ RTE Experiments
  - ❏ RTE Test Suite
  - ❏ Results
- ❏ Conclusion

RTE Experiment

- PASCAL RTE(3) (Recognising Textual Entailment) Dataset
  - Real ‘in the wild’ examples (FraCaS: textbook examples)
  - Longer examples; longer edits
  - No ‘unk’ -> ‘yes’ or ‘no’ (‘no’ includes ‘unk’)
  - More complicated inference
    - Temporal reasoning, relation extraction paraphrase
- Use of data
  - Subset of problems
  - Stanford RTE System
    - Use Stanford to align RTE data (Map H words to P words)
    - Can translate to NatLog alignment steps (insert, delete, substitute, advance)
    - Feed (translated) Stanford output into Entailment Classifier from NatLog

RTE 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: Mr. Fitzgerald revealed he was one of several top officials who told Mr. Libby in June 2003 that Valerie Plame, wife of the former ambassador Joseph Wilson, worked for the CIA.**
- h: Joseph Wilson worked for CIA.
- no

RTE Results

- High Precision compared to Stanford
  - positive predictions for 24% of cases
  - Bos and Markert had 4%
  - Similar precision though
- Hybrid System
  - Stanford gives yes/no threshold
  - +- x based on NatLog decision
  - "Balanced"
    - Best X value
    - Threshold to balance yes/no
  - "Optimized"
    - Threshold to get highest accuracy
    - Lots of "Yes" predictions

RTE3 Development Set (800 problems)				
System	% yes	precision	recall	accuracy
Stanford	50.25	68.66	66.99	67.25
NatLog	18.00	**76.39**	26.70	58.00
Hybrid, bal.	50.00	69.75	67.72	68.25
Hybrid, opt.	55.13	69.16	**74.03**	**69.63**

RTE3 Test Set (800 problems)				
System	% yes	precision	recall	accuracy
Stanford	50.00	61.75	60.24	60.50
NatLog	23.88	**68.06**	31.71	57.38
Hybrid, bal.	50.00	64.50	62.93	63.25
Hybrid, opt.	54.13	63.74	**67.32**	**63.62**

3.12% more accurate than Stanford ($p < 0.01$)
25 problems extra

RTE Results Discussion

- Natlog gives ‘no’ for first example (incorrect)
  - Can’t insert ‘called’
- Also gives ‘no’ for second example (correct)
  - Can’t insert ‘first’; Less precise Stanford is happy to add ‘first’ and gives yes.

ID	Premise(s)	Hypothesis	Answer
518	The French railway company SNCF is cooperating in the project.	The French railway company is called SNCF.	*yes*
601	NUCOR has pioneered a giant mini-mill in which steel is poured into continuous casting machines.	Nucor has pioneered the first mini-mill.	*no*

- Hybrid improved results aren’t because of monotonicity
  - Monotonicity is rare in real data
  - But NatLog is more precise

Outline

- Introduction
  - Shallow/Robust - Deep/Brittle tradeoff
  - Limits of Natural Logic
- Natural Logic Foundations
  - Entailment and Monotonicity
  - Composition
- NatLog System
  - Preprocessing, Alignment, Entailment classification
- Fracas Experiments
  - FraCaS test suite
  - Results
- RTE Experiments
  - RTE Test Suite
  - Results
- Conclusion

Conclusion

- Actually monotonicity is not very useful for textual entailment
  - But can be used to improve precision of existing systems

Questions?