Combining Natural Logic and Shallow Reasoning for Question Answering

Gabor Angeli, Neha Nayak, Chris Manning

Stanford University

August 8, 2016
Old Problem: Logic + ML often at odds
Old Problem: Logic + ML often at odds

ML gives us practical, generalizable systems:

P: *Ovaries are the female part of the flower, which produces eggs that are needed for making seeds.*

H: *A flower produces the seeds.*

...But struggles with logical subtleties

P: *Eating candy for dinner is an example of a poor health habit.*

H: *Eating candy is an example of a good health habit.*
Make ML more first-order-logic-like

Markov Logic Networks
- [Richardson and Domingos, 2006]
- [Niu et al., 2011]

Probabilistic Soft Logic
- [Kimmig et al., 2012]
- [Beltagy et al., 2014]

Deep Learning + Logic
- [Rocktäschel et al., 2014]
- [Bowman, 2014]
Find logics that are better for ML

Natural Logic!

The Persians are Invading Greece

All heroes are Persian!

Clearly you are wrong. You see, all Gods live on Olympus. Some heroes are Gods. And no one who lives on Olympus is Persian.
All heroes are Persian!

Clearly you are wrong. You see, all Gods live on Olympus. Some heroes are Gods. And no one who lives on Olympus is Persian.
### Show of Hands: First Order Logic?

1. $\forall x \ God(x) \supset LivesOnOlympus(x)$
2. $\exists x \ Hero(x) \land God(x)$
3. $\neg \exists x \ LivesOnOlympus(x) \land Persian(x)$

<table>
<thead>
<tr>
<th>Line</th>
<th>Rule</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$\forall x \ Hero(x) \supset Persian(x)$</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$a$</td>
<td>Hero($a$) \land God($a$)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Hero($a$) \land E, 5</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Hero($a$) \supset Persian($a$) \land E, 4</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Persian($a$) \land \Rightarrow E, 6, 7</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>God($a$) \land \land E, 5</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>God($a$) \supset LivesOnOlympus($a$) \land \land E, 1</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>LivesOnOlympus($a$) \land \Rightarrow E, 9, 10</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>LivesOnOlympus($a$) \land Persian($a$) \land \land I, 8, 11</td>
</tr>
<tr>
<td>13</td>
<td>$\exists x$</td>
<td>LivesOnOlympus($x$) \land Persian($x$) \land \exists I, 12</td>
</tr>
<tr>
<td>14</td>
<td>$\exists x$</td>
<td>LivesOnOlympus($x$) \land Persian($x$) \land R, 12</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>$\bot$ \land \land I, 3, 14</td>
</tr>
<tr>
<td>16</td>
<td>$\neg \forall x$</td>
<td>Hero($x$) \supset Persian($x$) \land \neg \land I, 4--15</td>
</tr>
</tbody>
</table>

---

Gabor Angeli, Neha Nayak, Chris Manning  
**Natural Logic + Shallow Reasoning**  
August 8, 2016 3 / 17
Syllogisms: The First Natural Logic

1. All Gods live on Mount Olympus
2. Some heroes are Gods
3. Nobody who lives on Mount Olympus is Persian
4. Some heroes live on Mount Olympus
5. Some heroes are not Persian
6. \( \neg \) All heroes are Persian

All (Darii), 1, 2
EIO (Ferio), 4, 3
SaP \( \perp \) SoP, 5
Syllogisms: The First Natural Logic

1. *All Gods live on Mount Olympus*
2. *Some heroes are Gods*
3. *Nobody who lives on Mount Olympus is Persian*
4. *Some heroes live on Mount Olympus*  \(\text{All (Darii), 1, 2}\)
5. *Some heroes are not Persian*  \(\text{EIO (Ferio), 4, 3}\)
6. \(\neg \text{ All heroes are Persian}\)  \(\text{SaP} \perp \text{SoP, 5}\)

...But syllogisms are crippling unexpressive
Quantifiers determine the *polarity* ($\uparrow$ or $\downarrow$) of words.

- **carnivores**
  - **felines**
    - house cats
  - **All**
- **consume**
  - **eat**
    - slurp
  - **mice**
    - fieldmice
- **placentals**
  - **rodents**
Quantifiers determine the *polarity* (↑ or ↓) of words. Mutations must respect *polarity*.
Quantifiers determines the *polarity* (↑ or ↓) of words. Mutations must respect *polarity*.
Quantifiers determines the *polarity* (↑ or ↓) of words. Mutations must respect *polarity*.

- **felines**: All ↑
  - **cats**: house cats ↓
  - **kitties**: consume ↑
- **rodents**: consume ↑
  - **mice**: fieldmice ↑
  - **pine vole**: eat
Quantifiers determine the *polarity* ($\uparrow$ or $\downarrow$) of words. Mutations must respect *polarity*. 

- **felines**: $\uparrow$ All $\downarrow$ house cats $\uparrow$ consume $\downarrow$ eat $\uparrow$ mice
- **placentals**: $\uparrow$ rods $\downarrow$ cats $\uparrow$ eat $\downarrow$ fieldmice

Gabor Angeli, Neha Nayak, Chris Manning
Quantifiers determine the *polarity* (↑ or ↓) of words. Mutations must respect *polarity*.

- **felines**: All house cats consumes mammals placentals
- **cats**: house cats eat rodents
- **kitties**: consume eat
- **mammals**: placentals

Gabor Angeli, Neha Nayak, Chris Manning (Stanford)
Quantifiers determines the *polarity* (↑ or ↓) of words.

Mutations must respect *polarity*.

Not pictured: also handles negation.
Shorthand for a node:

- **organism**
- **animal**
- **carnivores**
- **felines**
- **consume**
- **eat**
- **slurp**
- **living thing**
- **organism**
- **animals**
- **chordate**

No carnivores eat animals?
Natural Logic Q/A as Search

No carnivores eat animals?

The carnivores eat animals

The cat eats animals

The cat ate an animal

The cat ate a mouse

All cats have tails

All kittens are cute

Gabor Angeli, Neha Nayak, Chris Manning
No carnivores eat animals?

The carnivores eat animals

No animals eat animals
Natural Logic Q/A as Search

No carnivores eat animals

The carnivores eat animals?

The feline eats animals

All carnivores eat animals
The carnivores eat animals.

The feline eats animals?

The cat eats animals.

The cat eats chordate.
The feline eats animals

The cat eats animals?

The cat eats chordates

The kitty eats animals

\( \varphi \Rightarrow \psi \)

\( \varphi \not\Rightarrow \psi \)

\( \varphi \Rightarrow \neg \psi \)
The cat eats animals.

The cat eats chordates?

The cat eats mice.
The cat eats dogs.

\[ \varphi \Rightarrow \psi \]

\[ \varphi \not\Rightarrow \psi \]

\[ \varphi \Rightarrow \neg \psi \]

\[ \equiv \]

\[ \equiv \]

any
The cat eats chordates

The cat eats mice?

The cat ate a mouse

The kitty eats mice

\[ \phi \implies \psi \]

\[ \phi \not\implies \psi \]

\[ \phi \implies \neg \psi \]

...
Three Contributions for Generalizable Inference

1. Partial order over meronymy + relations

   Earth
   EU
   ↑ Germany
   Berlin
   have
   ↑ own
   sell

2. Natural Logic over dependency trees

   Some
   ↑ truly
   ↑ notorious
   ↑ villains
   ↑ have
   ↑ lairs
   operator
   nsubj
   amod
   advmod
dobj

3. Hybrid statistical / logical solver
Three Contributions for Generalizable Inference

1. Partial order over meronymy + relations

   Earth
   EU
   Germany
   have
   own
   Berlin

2. Natural Logic over dependency trees

   Some
   truly
   notorious
   villains
   have
   lairs

   operator
   advmod
   amod
   nsubj
   dobj
Three Contributions for Generalizable Inference

1. Partial order over meronymy + relations

Earth

EU

↑ Germany

have

Berlin

↑ own

sell

2. Natural Logic over dependency trees

Some↑ truly↑ notorious↑ villains↑ have↑ lairs↑.

Some

operator

advmod

amod

nsubj

dobj

3. Hybrid statistical / logical solver
Rain and snow are types of precipitation?

Rain and snow are forms of precipitation

Heavy rain and snow are types of precipitation
Rain and snow are types of precipitation?

Rain and snow are forms of precipitation.

Heavy rain and snow are types of precipitation.
Rain and snow are types of precipitation.

Rain and snow are forms of precipitation?

Heavy rain and snow are forms of precipitation.

Rain and snow are forms of weather.
Rain and snow are types of precipitation

Rain and snow are forms of precipitation?

Heavy rain and snow are forms of precipitation

Forms of precipitation include rain and sleet
Rain and snow are types of precipitation.

Rain and snow are forms of precipitation?

Rain and snow are forms of weather.

Heavy rain and snow are forms of precipitation.

Forms of precipitation include rain and sleet.
Forms of precipitation include rain and sleet.

Rain and snow are forms of precipitation.
Lexical Alignment Classifier

**Forms of precipitation include rain and sleet**

**Rain and snow are forms of precipitation**
Forms of precipitation include rain and sleet.

Rain and snow are forms of precipitation.
Forms of precipitation include rain and sleet.

Rain and snow are forms of precipitation.

Features

1. Matching words
**Lexical Alignment Classifier**

**Forms** of precipitation include **rain** and **sleet**

**Rain** and **snow** are **forms** of **precipitation**

**Features**

1. Matching words
2. Mismatched words
Forms of precipitation include rain and sleet.

Rain and snow are forms of precipitation.

Features

1. Matching words
2. Mismatched words
3. Unmatched words in premise/consequent
Lexical Alignment Classifier

**Forms of precipitation** include **rain** and **sleet**.

**Rain and snow** are **forms of precipitation**.

**Features**

1. Matching words
2. Mismatched words
3. Unmatched words in premise/consequent

**Competitive with Stanford RTE system (63% on RTE3)**
Old Problem: Logic + Lexical Classifiers

FOL and lexical classifiers don’t speak the same language
Old Problem: Logic + Lexical Classifiers

FOL and lexical classifiers don’t speak the same language
...but natural logic does!

I SPEAK WHALE!
Run our usual search

1. If we find a premise, great!
Big Picture

Run our usual search

1. If we find a premise, great!
2. If not, use lexical classifier as an evaluation function
Big Picture

Run our usual search

1. If we find a premise, great!
2. If not, use lexical classifier as an *evaluation function*

Visit 1M nodes / second: We have to be fast!
Dissecting Our Classifier

Anatomy of a Classifier

- Features $f$ (matching / mismatched / unmatched words)
- Weights $w$
- Entailment pair $x$

$$ p(\text{entail} \mid x) = \frac{1}{1+\exp(-w^T f(x))} $$
Dissecting Our Classifier

Anatomy of a Classifier

- Features $f$ (matching / mismatched / unmatched words)
- Weights $w$
- Entailment pair $x$

\[
p(\text{entail} \mid x) = \frac{1}{1 + \exp(-w^T f(x))}
\]

$p(\text{entail} \mid x)$ monotone w.r.t. $(w^T f(x))$
Dissecting Our Classifier

Anatomy of a Classifier

- Features $f$ (matching / mismatched / unmatched words)
- Weights $w$
- Entailment pair $x$

\[
p(\text{entail} \mid x) = \frac{1}{1 + \exp(-w^T f(x))}
\]

$p(\text{entail} \mid x)$ monotone w.r.t. $(w^T f(x))$

- Only need $w^T f(x)$ during search to compute $\max p(\text{entail} \mid x)$
- $w^T f(x)$ is our evaluation function
Incorporating our Evaluation Function

Anatomy of a Search Step

1. Mutate a word, or
2. Delete a word, or
3. Insert a word.

Each step updates a small number of features

\[ w^T f(x) = v \]
Incorporating our Evaluation Function

Anatomy of a Search Step

1. Mutate a word, or
2. Delete a word, or
3. Insert a word.

Each step updates a small number of features

\[ w^T f(x) = v \]
Incorporating our Evaluation Function

Anatomy of a Search Step

1. Mutate a word, or
2. Delete a word, or
3. Insert a word.

Each step updates a small number of features

\[ v' = v - w_i \cdot f_i + w_i \cdot f_i \]
Why is this Important?

Faster Search $\Rightarrow$ Deeper Reasoning

- **Speed**: Around 1M search states visited per second
- **Memory**: 32 byte search states

**Speed**: Don’t re-featurize at every timestep.

**Memory**: Never store intermediate fact as String.
Another Example Search

**Forms of precipitation include rain and sleet.**

**Rain and snow are types of precipitation.**

Score $w^T f(x): -0.5$

<table>
<thead>
<tr>
<th>Feature</th>
<th>$w$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching words</td>
<td>2.0</td>
<td>2</td>
</tr>
<tr>
<td>Mismatched words</td>
<td>-1.0</td>
<td>2</td>
</tr>
<tr>
<td>Unmatched premise</td>
<td>-0.5</td>
<td>1</td>
</tr>
<tr>
<td>Unmatched consequent</td>
<td>-0.75</td>
<td>0</td>
</tr>
<tr>
<td>Bias</td>
<td>-2.0</td>
<td>1</td>
</tr>
</tbody>
</table>
Rain and snow are **types of precipitation**?

Rain and snow are **forms of precipitation**

Rain and snow are **types of weather**
Another Example Search

**Forms of precipitation include** rain **and** sleet

Rain **and** snow are **forms of precipitation**

**Score** $w^T f(x): -0.5 + 2 - 1$

<table>
<thead>
<tr>
<th>Feature</th>
<th>$w$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching words</td>
<td>2.0</td>
<td>3</td>
</tr>
<tr>
<td>Mismatched words</td>
<td>-1.0</td>
<td>1</td>
</tr>
<tr>
<td>Unmatched premise</td>
<td>-0.5</td>
<td>1</td>
</tr>
<tr>
<td>Unmatched consequent</td>
<td>-0.75</td>
<td>0</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>-2.0</td>
<td>1</td>
</tr>
</tbody>
</table>
Another Example Search

**Forms of precipitation include** rain and sleet.

Rain and snow are forms of precipitation.

**Score** $w^T f(x): 2.5$

<table>
<thead>
<tr>
<th>Feature</th>
<th>$w$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching words</td>
<td>2.0</td>
<td>3</td>
</tr>
<tr>
<td>Mismatched words</td>
<td>-1.0</td>
<td>1</td>
</tr>
<tr>
<td>Unmatched premise</td>
<td>-0.5</td>
<td>1</td>
</tr>
<tr>
<td>Unmatched consequent</td>
<td>-0.75</td>
<td>0</td>
</tr>
<tr>
<td>Bias</td>
<td>-2.0</td>
<td>1</td>
</tr>
</tbody>
</table>
Which activity is an example of a good health habit?
(A) Watching television
(B) Smoking cigarettes
(C) Eating candy
(D) Exercising every day

Not smoking or drinking alcohol is good for health, regardless of whether clothing is worn or not.
Eating candy for dinner is an example of a poor health habit.
Multiple choice questions from real 4th grade science exams

Which activity is an example of a good health habit?
(A) Watching television
(B) Smoking cigarettes
(C) Eating candy
(D) Exercising every day
Solving 4th Grade Science

Multiple choice questions from real 4th grade science exams

Which activity is an example of a good health habit?
(A) Watching television
(B) Smoking cigarettes
(C) Eating candy
(D) Exercising every day

In our corpus:
- Plasma TV’s can display up to 16 million colors ... great for watching TV ... also make a good screen.
- Not smoking or drinking alcohol is good for health, regardless of whether clothing is worn or not.
- Eating candy for dinner is an example of a poor health habit.
- Healthy is exercising
Solving 4<sup>th</sup> Grade Science

Multiple choice questions from real 4<sup>th</sup> grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td>57</td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Hixon et al., 2015]
## Solving 4\textsuperscript{th} Grade Science

### Multiple choice questions from real 4\textsuperscript{th} grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td></td>
<td>57</td>
</tr>
<tr>
<td>IR Baseline</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>This Work</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

We're able to pass 4\textsuperscript{th} grade science! (Though AI2’s recent systems outperform this)

[Hixon et al., 2015]
Solving 4\textsuperscript{th} Grade Science

Multiple choice questions from real 4\textsuperscript{th} grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>IR Baseline</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>This Work</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>More Data + IR Baseline</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>More Data + This Work</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

[We're able to pass 4\textsuperscript{th} grade science! (Though AI2's recent systems outperform this)]

[Hixon et al., 2015]
Solving 4th Grade Science

Multiple choice questions from real 4th grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td>57</td>
<td>–</td>
</tr>
<tr>
<td>IR Baseline</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>This Work</td>
<td>52</td>
<td>51</td>
</tr>
<tr>
<td>More Data + IR Baseline</td>
<td>62</td>
<td>58</td>
</tr>
<tr>
<td>More Data + This Work</td>
<td>65</td>
<td>61</td>
</tr>
</tbody>
</table>

[We're able to pass 4th grade science! (Though AI2's recent systems outperform this)]

[Hixon et al., 2015]
## Solving 4th Grade Science

### Multiple choice questions from real 4th grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td>57</td>
<td>–</td>
</tr>
<tr>
<td>IR Baseline</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>This Work</td>
<td>52</td>
<td>51</td>
</tr>
<tr>
<td>More Data + IR Baseline</td>
<td>62</td>
<td>58</td>
</tr>
<tr>
<td>More Data + This Work</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>This Work + 📣 + 🎯</td>
<td>74</td>
<td>67</td>
</tr>
</tbody>
</table>

[Hixon et al., 2015]

Though AI2’s recent systems outperform this.
Solving 4\textsuperscript{th} Grade Science

Multiple choice questions from real 4\textsuperscript{th} grade science exams

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWBOT</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>KNOWBOT (ORACLE)</td>
<td>57</td>
<td>–</td>
</tr>
<tr>
<td>IR Baseline</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>This Work</td>
<td>52</td>
<td>51</td>
</tr>
<tr>
<td>More Data + IR Baseline</td>
<td>62</td>
<td>58</td>
</tr>
<tr>
<td>More Data + This Work</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>This Work + 🔔 + 🔔</td>
<td>74</td>
<td>67</td>
</tr>
</tbody>
</table>

We’re able to pass 4\textsuperscript{th} grade science!  
(Though AI2’s recent systems outperform this)

[Hixon et al., 2015]
Conclusions

Natural Logic

- A logic over the syntax of natural language
- Expressive but efficient

Natural Logic “plays nice” with statistical (/deep?) methods

- Both operate directly over text
- Use lexical classifier as evaluation function

NaturalLI + Evaluation Function

- Also detects *likely* entailment / contradictions
- 3% improvement on science exam questions
References I


Tuffy: Scaling up statistical inference in markov logic networks using an rdbms.
*VLDB.*

Markov logic networks.

Low-dimensional embeddings of logic.
In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*.

*Studies on natural logic and categorial grammar.*
PhD thesis, University of Amsterdam.