An Inference-Based Approach to Recognizing Entailment

Peter Clark and Phil Harrison
Boeing Research and Technology
Seattle, WA
Outline

- **BLUE** (our system)
  - Description
  - Good and bad examples on RTE5
  - Performance and ablations on RTE5
- **Reflections**
  - The Knowledge Problem
  - The Reasoning Problem
- Parse, *generate logic* for T and H
- See if every *clause* in H *subsumes* part of T
- Use DIRT and WordNet
T: A black cat ate a mouse.
H: A mouse was eaten by an animal.

modifier(cat01,black01),
subject(eat01,cat01),
object(eat01,mouse01).

subject(eat01,animal01),
object(eat01,mouse01).

subsumes?
1. The Logic Module: Generating a Representation

"A black cat ate a mouse."

```
(DECL ((VAR _X1 "a" "cat" (AN "black" "cat")))
     (VAR _X2 "a" "mouse"))
(S (PAST) _X1 "eat" _X2))
```

```
"cat"(cat01),
"black"(black01),
"eat"(eat01),
"mouse"(mouse01),
modifier(cat01,black01),
subject(eat01,cat01),
object(eat01,mouse01).
```

```
cat#n1(cat01),
black#a1(black01),
mouse#n1(mouse01),
eat#v1(eat01),
color(cat01,black01),
agent(eat01,cat01),
object(eat01,mouse01).
```
1. The Logic Module: Lexico-Semantic Inference

- Computing subsumption (= entailment)

\[ T: \text{A black cat ate a mouse} \]
\[ \text{subject(eat01,cat01), object(eat01,mouse01), mod(cat01,black01)} \]

\[ \text{“by”(eat01,animal01), object(eat01,mouse01)} \]

\[ H: \text{A mouse was eaten by an animal} \]
1. The Logic Module: Lexico-Semantic Inference

- Subsumption

**T:** A black cat ate a mouse
subject(eat01,cat01), object(eat01,mouse01), mod(cat01,black01)

“by”(eat01,animal01), object(eat01,mouse01)

**H:** A mouse was eaten by an animal

WordNet

- cat#n1 hypernym animal#n1
- speedy#s2 similar-to fast#a1
- rapidly#r1 pertains-to quick#a1
- destroy#v1 derives destruction#n1
T: A black cat ate a mouse

If X eats Y then X chews Y

If X eats Y then X digests Y

T': A black cat ate a mouse. The cat is black. The cat digests the mouse. The cat chewed the mouse. The cat swallows the mouse...
With Inference...

T: A black cat ate a mouse

IF X eats Y THEN X chews Y

IF X eats Y THEN X digests Y

T': A black cat ate a mouse. The cat is black. The cat digests the mouse. The cat chewed the mouse. The cat swallows the mouse...

H: An animal digested the mouse.

H entailed!
Ignore syntactic structure:

- Use bag of words for T and H
- See if every word in H subsumes one in T
- Use DIRT and WordNet
T: A black cat ate a mouse.
H: A mouse was digested by an animal.

T,H → Logic Representation

YES/NO → YES

T,H → Bag-of-Words Representation

UNKNOWN → UNKNOWN

WordNet
DIRT

SUBSUMES?

{ black cat eat mouse }

{ mouse digest animal }
Bag of Words Inference

T: A black cat ate a mouse.
H: A mouse was digested by an animal.

\[
\begin{align*}
T &\quad \{ \text{black cat eat mouse} \} \\
H &\quad \{ \text{mouse digest animal} \}
\end{align*}
\]
Bag of Words Inference

T: A black cat ate a mouse.
H: A mouse was digested by an animal.

\[
\begin{align*}
T & \quad \{ \text{black cat eat mouse} \} \\
H & \quad \{ \text{mouse digest animal} \} \\
\text{WordNet} & \quad \text{cat#n1 hypernym animal#n1}
\end{align*}
\]
Bag of Words Inference

T: A black cat ate a mouse.
H: A mouse was digested by an animal.

T
{ black cat eat mouse }

H
{ mouse digest animal }

DIRT
IF X eats Y THEN X digests Y

“eat” → “digest”
Bag of Words Inference

T: A black cat ate a mouse.
H: A mouse was digested by an animal.

T: { black cat eat mouse }
H: { mouse digest animal }

H entailed!
Outline

- **BLUE** (our system)
  - Description
  - Good and bad examples on RTE5
  - Performance and ablations on RTE5
- **Reflections**
  - The Knowledge Problem
  - The Reasoning Problem
The Good…

#191 (BLUE got this right)

T: …Ernie Barnes…was an offensive linesman….
H: Ernie Barnes was an athlete.

via WordNet: linesman#n1 isa athlete#n1

#333 (BLUE got this right)

T: …hijacking of a Norwegian tanker…by Somali pirates
H: Somali pirates attacked a Norwegian tanker.

via DIRT: IF X hijacks Y THEN Y is attacked by X.

Pilot H26 (BLUE got this right)

T: …Charles divorced Diana…
H: Prince Charles was married to Princess Diana.

via DIRT: IF X divorces Y THEN X marries Y.
T: Crippa died...after he ate...deadly...wild mushrooms
H: Crippa was killed by a wild mushroom.

via DIRT: IF X dies of Y THEN Y is killed by X

Pilot H142 (BLUE got this right)

HEADLINE: EU slams Nepalese king's dismissal...
T: The EU...presidency called for ...democracy.
H: There has been a...call for ..democracy in Nepal

via use of HEADLINE as context (and WordNet Nepalese/Nepal)
### The Bad

**#407 (BLUE got this wrong, predicting YES)**

T: Venus Williams *triumphed* over…Bartoli…

H*: Venus Williams was *defeated by*…Bartoli…

via (bad rule in) DIRT: IF Y wins over X THEN X defeats Y.

---

**#219 (BLUE got this right, but for nonsensical reasons)**

T: PepsiCo …acquired …Star Foods…

H: PepsiCo holds Star Foods

via DIRT: IF X acquires Y THEN X sells Y

and: IF Y sells X’s business THEN Y holds X’s tongue

and WordNet: “tongue” isa “food”

---

**Pilot H29 (BLUE got this wrong, predicting YES)**

T: …even if Iceland offered Fischer citizenship….

H*: Iceland granted Bobby Fischer citizenship.

BLUE does not recognize the hypothetical blocks entailment
T: …Slumdog Millionaire director Danny Boyle….
H: The *movie* "Slumdog Millionaire" has been directed by Danny Boyle.

*(unable to conclude “movie” in H)*

Pilot H75 (BLUE got this wrong, predicting YES)

T: ..the oath taken by the 115 electors…
H*: The *cardinals* electing the pope…

“115 is a cardinal” (!)
Outline

- **BLUE** (our system)
  - Description
  - Good and bad examples on RTE5
  - Performance and ablations on RTE5
- **Reflections**
  - The Knowledge Problem
  - The Reasoning Problem
## Results: Main Task

<table>
<thead>
<tr>
<th></th>
<th>2-Way</th>
<th>3-way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic module only</td>
<td>56.7</td>
<td>46.3</td>
</tr>
<tr>
<td>Bag-of-words only</td>
<td>60.0</td>
<td>52.8</td>
</tr>
<tr>
<td>Logic + bag-of-words</td>
<td>61.5</td>
<td>54.7</td>
</tr>
</tbody>
</table>

- Pipeline works best
## Results: Main Task

<table>
<thead>
<tr>
<th></th>
<th>2-Way</th>
<th>3-way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic module only</td>
<td>56.7</td>
<td>46.3</td>
</tr>
<tr>
<td>Bag-of-words only</td>
<td>60.0</td>
<td>52.8</td>
</tr>
<tr>
<td>Logic + bag-of-words</td>
<td>61.5</td>
<td>54.7</td>
</tr>
</tbody>
</table>

- Pipeline works best
- Logic alone is worse than bag alone
  - Only decides 29% of cases, but does well (64%) on these
## Ablation Studies

<table>
<thead>
<tr>
<th></th>
<th>RTE5 Dev</th>
<th>RTE5 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full BLUE</td>
<td>63.8</td>
<td>61.5</td>
</tr>
<tr>
<td>- remove WordNet</td>
<td>- 6.0</td>
<td>- 4.0</td>
</tr>
<tr>
<td>- remove DIRT</td>
<td>- 0.5</td>
<td>+1.2</td>
</tr>
<tr>
<td>- remove parsing</td>
<td>- 0.3</td>
<td>- 1.5</td>
</tr>
</tbody>
</table>
## Ablation Studies

<table>
<thead>
<tr>
<th></th>
<th>RTE5 Dev</th>
<th>RTE5 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full BLUE</strong></td>
<td>63.8</td>
<td>61.5</td>
</tr>
<tr>
<td>- remove WordNet</td>
<td>- 6.0</td>
<td>- 4.0</td>
</tr>
<tr>
<td>- remove DIRT</td>
<td>- 0.5</td>
<td>+1.2</td>
</tr>
<tr>
<td>- remove parsing</td>
<td>- 0.3</td>
<td>- 1.5</td>
</tr>
</tbody>
</table>
Ablation Studies

<table>
<thead>
<tr>
<th></th>
<th>RTE5 Dev</th>
<th>RTE5 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full BLUE</td>
<td>63.8</td>
<td>61.5</td>
</tr>
<tr>
<td>- remove WordNet</td>
<td>-6.0</td>
<td>-4.0</td>
</tr>
<tr>
<td>- remove DIRT</td>
<td>-0.5</td>
<td>+1.2</td>
</tr>
<tr>
<td>- remove parsing</td>
<td>-0.3</td>
<td>-1.5</td>
</tr>
</tbody>
</table>

- WordNet is significantly helping
**Ablation Studies**

<table>
<thead>
<tr>
<th></th>
<th>RTE5 Dev</th>
<th>RTE5 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full BLUE</td>
<td>63.8</td>
<td>61.5</td>
</tr>
<tr>
<td>- remove WordNet</td>
<td>- 6.0</td>
<td>- 4.0</td>
</tr>
<tr>
<td>- remove DIRT</td>
<td>- 0.5</td>
<td>+1.2</td>
</tr>
<tr>
<td>- remove parsing</td>
<td>- 0.3</td>
<td>- 1.5</td>
</tr>
</tbody>
</table>

- WordNet is significantly helping
- DIRT is barely helping
  - **Rules are noisy** (≈ 50% are bad)
  - **Applicability is low** (≈ 10%-15%) – most RTE problems are outside DIRT’s scope
Ablation Studies

<table>
<thead>
<tr>
<th></th>
<th>RTE5 Dev</th>
<th>RTE5 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full BLUE</td>
<td>63.8</td>
<td>61.5</td>
</tr>
<tr>
<td>- remove WordNet</td>
<td>- 6.0</td>
<td>- 4.0</td>
</tr>
<tr>
<td>- remove DIRT</td>
<td>- 0.5</td>
<td>+1.2</td>
</tr>
<tr>
<td>- remove parsing</td>
<td>- 0.3</td>
<td>- 1.5</td>
</tr>
</tbody>
</table>

- WordNet is significantly helping
- DIRT is barely helping
- Parsing is barely helping
  - Extracting syntactic structure is **very error-prone**
  - **Semantic relationships usually persist** from T to H and H*
    - Non-entailment caused by *other* factors
How important is semantic (hence syntactic) structure?

T: Boyle \textbf{directed} Slumdog Millionaire
H*: Slumdog Millionaire \textbf{directed} Boyle [NOT entailed]

but this kind of example is unusual in RTE!

\textbf{“Semantic Continuity” Conjecture:}

IF T and H are \textit{“sensible”}
AND T and H are consistent with world knowledge
AND T and H are \textit{topically similar}
THEN this \textbf{heavily constrains} the variability in possible semantic (hence syntactic) relationships

\textarrow{reduced discriminatory power} of semantic/syntactic analysis
Outline

- BLUE (our system)
  - Description
  - Good and bad examples on RTE5
  - Performance and ablations on RTE5
- Reflections
  - The Knowledge Problem
  - The Reasoning Problem
What are the Long Term Challenges?

- The Knowledge Problem
  - Still missing a *lot* of world knowledge

T: ... volunteers...helped...**create dikes** to protect... against the Red River...**flood**.
H: Red River will **overflow** its **banks**.

- Need to know...
  - **flood**: water...overflowing onto normally dry land
  - **bank**: sloping land..beside..water
What are the Long Term Challenges?

- The Knowledge Problem
  - Still missing a *lot* of world knowledge

#341

T: ... volunteers...helped...**create dikes** to protect... against the Red River...**flood**.
H: Red River will **overflow** its **banks**.

- Need to know...
  - **flood**: water...overflowing onto normally dry land
  - **bank**: sloping land..beside..water

**includes**
What are the Long Term Challenges?

- The Knowledge Problem
  - Still missing a *lot* of world knowledge
- The Reasoning Problem
  - Finding *some* path from T to H is error-prone
What are the Long Term Challenges?

- **The Knowledge Problem**
  - Still missing a *lot* of world knowledge

- **The Reasoning Problem**
  - Finding *some* path from T to H is error-prone

---

T: Venus Williams *triumphed* over Bartoli…to win…
H*: Venus Williams was *defeated by*…Bartoli…

IF Y triumphs over X THEN X defeats Y  \[\rightarrow\] Wrong 😞

- **BUT**: evidence *against* H:
  - triumph=defeat, and defeat is antisymmetric
  - World Knowledge: “win” implies defeat (not defeated by)

- **Better**: look at multiple reasoning paths
  find the “best”, consistent subset of implications
Williams triumphed over Bartoli.
Venus Williams triumphed over Bartoli to win...
Venus Williams triumphed over Bartoli to win…

Was Williams defeated? **Answer: No!**

What is the overall scene?
Williams triumphed over Bartoli to win...

**“T” text:**
- Williams triumphed over Bartoli
- Williams won

**“H” text:**
- Williams was defeated by Bartoli
- Williams triumphed
- Williams lost to Bartoli
- Williams had a victory
- Williams defeated someone

Williams was defeated? **Answer: No!**

What is the overall scene? **Answer:**
Venus Williams triumphed over Bartoli to win...
Summary

- **BLUE:**
  - Pipeline of *logical representation + bag-of-words*
  - Reasoning with *WordNet* and *DIRT*
  - Performance – ok (above the median)

- **Ablations:**
  - WordNet helps a lot
  - DIRT and parsing barely helped

- Two big challenges:
  - **Knowledge** – need lots more
  - **Reasoning** – need search for coherence, not a single path

Thank you!