Johan Bos & Katja Markert (2005): Recognizing Textual Entailment with Logical Inference

Feraena Bibyna

January 28, 2014
Recognizing Textual Entailment (RTE)

Does \( T \) entail \( H \)?

\( T: \) In 1998, the General Assembly of the Nippon Sei Ko Kai (Anglican Church in Japan) voted to accept female priests.

\( H: \) The Anglican church in Japan approved the ordination of women.

\( T: \) The city Tenochtitlan grew rapidly and was the center of the Aztec’s great empire.

\( H: \) Tenochtitlan quickly spread over the island, marshes, and swamps.

**Ultimate** challenges for NLP systems!
This Paper

Two methods for RTE:

- **Shallow method**: word overlap
- **Deep semantic analysis**: theorem prover, model builder

Research question:

- Improvement over baseline? Hybrid system over individual use?
- Effect of lack of lexical & world knowledge on deep semantic analysis? How can we do logical inference despite of this?
- Effect of test suite on performance?
Shallow Semantic Features

Expect some dependency between surface string similarity and the existence of entailment

Lemma $l_1$ and $l_2$ is related iff $l_1$ and $l_2$...

- ... are equal
- ... belong to the same WordNet synset (murder and slay)
- ... are related via WordNet derivation (murder and murderer)
- ... are related via a combination of synonymy and derivations (murder and liquidator via murderer)
Each lemma in hypothesis is assigned its inverse document frequency as its **weight**

\[
wn_{overlap} = \frac{\sum \text{weight of lemmas in } H \text{ related to lemmas in } T}{\sum \text{all weight of lemmas in } H}
\]

Taking into account length of text (in true entailments hypothesis is shorter than the text)
Deep Semantic Analysis

- CCG-parser (Bos et al., 2004)
- Semantic is represented by first-order fragment of DRS-language used in Discourse Representation Theory (Kamp & Reyle, 1993)
- Entailment checking:
  - **Vampire**, a theorem prover (Riazanov & Voronkov, 2002)
  - **Paradox**, a model builder (Claessen & Sörrensson, 2003)
Example: 78 (FALSE)

\[-\text{T: Clinton's new book is not big seller here.}\]
\[-\text{H: Clinton's book is a big seller.}\]

\[
\begin{array}{c}
\text{drs(T):} \\
\text{book(x1)} \\
\text{book(x2)} \\
\text{\neg x1=x2} \\
\text{clinton(x3)} \\
\text{of(x1,x3)} \\
\text{e4 x5} \\
\text{\neg big(x5)} \\
\text{seller(x5)} \\
\text{be(e4)} \\
\text{agent(e4,x1)} \\
\text{patient(e4,x5)} \\
\text{loc(e4,here)} \\
\end{array}
\]

\[
\begin{array}{c}
\text{drs(H):} \\
\text{book(x1)} \\
\text{clinton(x2)} \\
\text{of(x1,x2)} \\
\text{big(x4)} \\
\text{seller(x4)} \\
\text{be(e3)} \\
\text{agent(e3,x1)} \\
\text{patient(e3,x4)} \\
\end{array}
\]
"A woman snorts."

$$
\exists x (\text{woman}(x) \land \text{snort}(x))
$$
Given a T/H pair, a theorem prover can be used to find answers for:

1. T implies H \( \Rightarrow \) \( \text{FOL}(\text{DRS}(T)) \rightarrow \text{FOL}(\text{DRS}(H)) \)
2. T + H are inconsistent \( \neg \) \( \neg \text{FOL}(\text{DRS}(T);\text{DRS}(H)) \)

Proving (1) would shows entailment, and proving (2) would shows no entailment
Background Knowledge

Example: 1952 (TRUE)

\[ T: \text{Crude oil prices soared to record levels.} \]
\[ H: \text{Crude oil prices rise.} \]

- Need to know that **soaring** is a way of **rising** \( \rightarrow \) **background knowledge**!
- Supply \( BK \land \text{FOL}(\text{drs}(T);\text{drs}(H)) \) to theorem prover
- Generic knowledge, lexical knowledge, and geographical knowledge
Model Building

- Theorem prover are not good at deciding that a formula is not a theorem.
- Model builders: show that formula is true in at least one model.
- Use both: theorem prover attempts to prove the input while model builder simultaneously tries to find a model for its negation.
Model Builder

D={d₁,d₂,d₃}

F(mia)=d₁
F(butch)=d₂
F(vincent)=d₃

F(man)={d₂,d₃}
F(woman)={d₁}
F(know)={((d₂,d₃),(d₃,d₁),(d₃,d₂)}

¬know(mia,vincent) ♦ ∀x(man(x) → ∃y(woman(y) ∧ know(x,y)))
man(butch) ♦ man(vincent) ♦ woman(mia) ♦ ∀x(man(x) → ¬woman(x))

D={d₁,d₂} F(mia)=d₁
F(butch)=d₂
F(vincent)=d₂

F(man)={d₂}
F(woman)={d₁}
F(know)={((d₂,d₁)}
...outputs a **model** \(< D, F >\) for its input formula.

\[
\text{T: Clinton’s new book is not big seller here.}
\]

\[
D = \{d1,d2,d3\} \quad F(\text{loc}) = \{}
\]
\[
F(\text{book}) = \{d1,d2\} \quad F(\text{seller}) = \{}
\]
\[
F(\text{clinton}) = \{d3\} \quad F(\text{be}) = \{}
\]
\[
F(\text{of}) = \{(d1,d3)\} \quad F(\text{agent}) = \{}
\]
\[
F(\text{big}) = \{} \quad F(\text{patient}) = \{}
\]
Approximating Entailment

- It’s extremely hard to acquire all the required knowledge
- Use the models produced by the model builders to measure "distance" from an entailment
- If H is entailed by T, the model for T+H wouldn’t introduce many new entities
  → **domain size** of T+H would be similar to domain size of T
Domain and Model Size

- Domain size of $\text{fol}(\text{drs}(T)) = 11$, $\text{fol}(\text{drs}(T);\text{drs}(H)) = 12$ $\rightarrow$ $T$ likely entails $T$

- Model size: number of all instances of two/three places relations in the model, multiplied by the domain size.

Example: 1049 (TRUE)

$T$: Four Venezuelan firefighters who were traveling to a training course in Texas were killed when their sport utility vehicle drifted onto the shoulder of a highway and struck a parked truck.

$H$: Four firefighters were killed in a car accident.
Domain and Model Size (2)

\[ D = \{d1, d2, d3\} \]
\[ F(\text{cat}) = \{d1, d2\} \]
\[ F(\text{john}) = \{d3\} \]
\[ F(\text{of}) = \{(d1,d3)\} \]
\[ F(\text{like}) = \{(d3,d1), (d3,d2)\} \]

Domain size = 3
Model size = 3 \times 3 = 9
Features relevant for recognizing textual entailment:

- **Theorem prover**: entailed, inconsistent
- **Model builder**:
  - domain-size, model-size
  - domain-size-abs-dif, model-size-abs-dif
  - domain-size-rel-dif, model-size-rel-dif
Dataset Design

- Test set with 50% TRUE, 50% FALSE
- Task variable:
  - Comparable Document (CD)
  - Question Answering (QA)
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Reading Comprehension (RC)
  - Paraphrase Acquisition (PP)
  - Information Retrieval (IR)
- Cover wide variety of different aspects of entailment
Evaluation Measures

- Expressed as feature vectors, then trained a decision tree for TRUE/FALSE classification using Weka (also computes confidence value)

- Evaluation measures:
  - accuracy \((\text{acc})\)
  - confidence-weighted average score \((cws)\)
  \[
cws = \frac{1}{n} \sum_{i=1}^{n} \frac{\# \text{ correct-up-rank-}i}{i}
\]
Experiment 1: Human Upper Bound

- Manually annotated by the one of the author
- Accuracy is compared to organizer’s gold standard annotation

<table>
<thead>
<tr>
<th>Task</th>
<th>acc</th>
<th>cws</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>0.967</td>
<td>n/a</td>
</tr>
<tr>
<td>IE</td>
<td>0.975</td>
<td>n/a</td>
</tr>
<tr>
<td>MT</td>
<td>0.900</td>
<td>n/a</td>
</tr>
<tr>
<td>QA</td>
<td>0.961</td>
<td>n/a</td>
</tr>
<tr>
<td>RC</td>
<td>0.979</td>
<td>n/a</td>
</tr>
<tr>
<td>PP</td>
<td>0.920</td>
<td>n/a</td>
</tr>
<tr>
<td>IR</td>
<td>0.922</td>
<td>n/a</td>
</tr>
<tr>
<td>all</td>
<td>0.951</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Experiment 2: Shallow Features

- High-performance on CD
- Overestimates the number of true entailment
  - TRUE: 0.926 recall, 0.547 precision
  - FALSE: 0.236 recall, 0.761 precision

... high overlap is **not sufficient** for true entailment
Experiment 3: Strict Entailment

- Only use entailment and inconsistent feature
- For TRUE class: 0.767 precision, 0.065 recall
- Overestimates the number of false entailment

...missing lexical and background knowledge
Experiment 4: Approximating Entailment

---

<table>
<thead>
<tr>
<th>Exp</th>
<th>3: Strict</th>
<th>4: Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>acc</td>
<td>cws</td>
</tr>
<tr>
<td>CD</td>
<td>0.547</td>
<td>0.617</td>
</tr>
<tr>
<td>IE</td>
<td>0.542</td>
<td>0.622</td>
</tr>
<tr>
<td>MT</td>
<td>0.500</td>
<td>0.436</td>
</tr>
<tr>
<td>QA</td>
<td>0.461</td>
<td>0.422</td>
</tr>
<tr>
<td>RC</td>
<td>0.557</td>
<td>0.638</td>
</tr>
<tr>
<td>PP</td>
<td>0.540</td>
<td>0.581</td>
</tr>
<tr>
<td>IR</td>
<td>0.489</td>
<td>0.421</td>
</tr>
<tr>
<td>all</td>
<td>0.520</td>
<td>0.548</td>
</tr>
</tbody>
</table>

Example: 1049 (TRUE)

\[ T: \text{Four Venezuelan firefighters who were traveling to a training course in Texas were killed when their sport utility vehicle drifted onto the shoulder of a highway and struck a parked truck.} \]

\[ H: \text{Four firefighters were killed in a car accident.} \]

...has similar result to shallow classifier, but shows more promising performance for several subsets
Experiment 5: Hybrid Classification

- Compared to shallow classifier: performs better or equally on all subsets but CD.
- Compared to deep classifier: performs better or equally on all subsets but MT.
Experiment 6: Dependency on Dataset Design

- Integrated the subset indicator as a feature.
- Using both a combination methodologies and the subset indicator is necessary to improve on individual shallow and deep classifiers.
Theorem proving is **not enough**.
- High precision, but low recall.
- Used model building to surmount this problem to a certain extent.
- Need incorporation of automatic methods for knowledge acquisition.

Hybrid approach achieves high accuracy.
- choice of entailment methods might have to vary according to dataset design/application
- integration of several entailment methods and indicator of design methodology are needed to achieve robust performance.