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**)

Shallow Semantic Features: Measures

- Each lemma in hypothesis is assigned its inverse document frequency as its **weight**
- $wnoverlap = \frac{\sum \text{weight of lemmas in H related to lemmas in T}}{\sum \text{all weight of lemmas in H}}$
- Taking into account length of text (in true entailments hypothesis is shorter than the text)

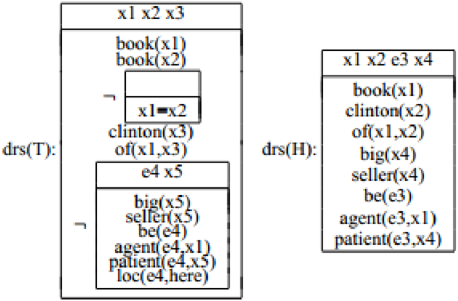
Deep Semantic Analysis

- 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)

Semantic Representation

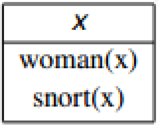
Example: 78 (FALSE)

T: Clinton's new book is not big seller here.
H: Clinton's book is a big seller.



DRS to FOL

"A woman snorts."



$\exists x(woman(x) \wedge snort(x))$

Theorem Prover

Given a T/H pair, a theorem prover can be used to find answers for:

1. T implies H $\quad \textsc{fol}(\textsc{drs}(\mathrm{T})) \rightarrow \textsc{fol}(\textsc{drs}(\mathrm{H}))$
2. T+H are inconsistent $\quad \neg\textsc{fol}(\textsc{drs}(\mathrm{T});\textsc{drs}(\mathrm{H}))$

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

Background Knowledge

Example: 1952 (TRUE)

T: Crude oil prices soared to record levels.
H: Crude oil prices rise.

- Need to know that **soaring** is a way of **rising** → **background knowledge**!
- Supply $\mathrm{BK} \wedge \textsc{fol}(\textsc{drs}(\mathrm{T});\textsc{drs}(\mathrm{H}))$ 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(man)={d2,d3}
F(butch)=d2          F(woman)={d1}
F(vincent)=d3        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(man)={d2}
          F(butch)=d2        F(woman)={d1}
          F(vincent)=d2      F(know)={(d2,d1)}
```

Model Building

...outputs a **model** ($< D, F >$) for its input formula.

T: Clinton's new book is not big seller here.

```
D = {d1,d2,d3}        F(loc) = {}
F(book) = {d1,d2}     F(seller) = {}
F(clinton) = {d3}     F(be) = {}
F(of) = {(d1,d3)}     F(agent) = {}
F(big) = {}           F(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
  $\rightarrow$ **domain size** of T+H would be similar to domain size of T

Domain and Model 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 size of fol(drs(T)) = 11, fol(drs(T);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.

Domain and Model Size (2)

```
D = {d1,d2,d3}
F(cat) = {d1,d2}
F(john) = {d3}
F(of) = {(d1,d3)}
F(like) = {(d3,d1),(d3,d2)}
```

Domain size = 3
Model size = 3 * 3 = 9

Deep Semantic Features

Features relevant for recognizing textual entailment:

- Theorem prover: `entailed`, `inconsistent`
- Model builder:
  - `domainsize`, `modelsize`
  - `domainsizeabsdif`, `modelsizeabsdif`
  - `domainsizereldif`, `modelsizereldif`

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 (*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

Exp	1: Human	
Task	acc	cws
CD	0.967	n/a
IE	0.975	n/a
MT	0.900	n/a
QA	0.961	n/a
RC	0.979	n/a
PP	0.920	n/a
IR	0.922	n/a
all	0.951	n/a

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

Experiment 2: Shallow Features

Exp	2: Shallow	
Task	acc	cws
CD	**0.827**	**0.881**
IE	0.508	0.503
MT	0.500	0.515
QA	0.531	0.557
RC	0.507	0.502
PP	0.480	0.467
IR	0.511	0.561
all	0.569	0.624

- 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

Exp	2: Shallow		3: Strict	
Task	acc	cws	acc	cws
CD	**0.827**	**0.881**	0.547	0.617
IE	0.508	0.503	**0.542**	0.622
MT	0.500	0.515	0.500	0.436
QA	0.531	0.557	0.461	0.422
RC	0.507	0.502	**0.557**	0.638
PP	0.480	0.467	0.540	0.581
IR	0.511	0.561	0.489	0.421
all	0.569	0.624	0.520	0.548

- 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

Exp	3: Strict		4: Deep	
Task	acc	cws	acc	cws
CD	0.547	0.617	0.713	0.787
IE	**0.542**	0.622	0.533	0.616
MT	0.500	0.436	**0.592**	**0.596**
QA	0.461	0.422	0.515	0.419
RC	**0.557**	0.638	0.457	0.537
PP	0.540	0.581	0.520	0.616
IR	0.489	0.421	0.567	0.503
all	0.520	0.548	0.562	0.608

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.

...has similar result to shallow classifier, but shows more promising performance for several subsets

Experiment 5: Hybrid Classification

Exp	2: Shallow		4: Deep		5: Hybrid	
Task	acc	cws	acc	cws	acc	cws
CD	**0.827**	**0.881**	0.713	0.787	0.700	0.790
IE	0.508	0.503	0.533	0.616	**0.542**	0.639
MT	0.500	0.515	**0.592**	**0.596**	0.525	0.512
QA	0.531	0.557	0.515	0.419	0.569	0.520
RC	0.507	0.502	0.457	0.537	0.507	0.587
PP	0.480	0.467	0.520	0.616	0.560	**0.667**
IR	0.511	0.561	0.567	0.503	**0.622**	**0.569**
all	0.569	0.624	0.562	0.608	0.577	0.632

- 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

Exp	2: Shallow		4: Deep		5: Hybrid		6: Hybrid+Task	
Task	acc	cws	acc	cws	acc	cws	acc	cws
CD	**0.827**	**0.881**	0.713	0.787	0.700	0.790	**0.827**	0.827
IE	0.508	0.503	0.533	0.616	**0.542**	0.639	**0.542**	**0.627**
MT	0.500	0.515	**0.592**	**0.596**	0.525	0.512	0.533	0.581
QA	0.531	0.557	0.515	0.419	0.569	0.520	**0.577**	**0.531**
RC	0.507	0.502	0.457	0.537	0.507	0.587	**0.557**	**0.644**
PP	0.480	0.467	0.520	0.616	0.560	**0.667**	**0.580**	0.619
IR	0.511	0.561	0.567	0.503	**0.622**	**0.569**	0.611	0.561
all	0.569	0.624	0.562	0.608	0.577	0.632	**0.612**	**0.646**

- 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.

Conclusion

- 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.