A Natural Proof System for Natural Language

NPS4NL-5: Natural Language Inference with Natural Theorem Prover

Lasha Abzianidze  Reinhard Muskens

ESSLLI 2019 in Rīga, Latvija
Today:

- Relevant NLI datasets:
  - FraCaS
  - SICK

- Learning phase:
  - Adaptation
  - Development

- Evaluation:
  - FraCaS
  - SICK

- Demo of LangPro

- Conclusion & future work
The SICK dataset

SICK [Marelli et al., 2014b] contains Sentences Involving Compositional Knowledge:

- 10K Text-Hypothesis pairs generated semi-automatically and annotated by humans with three labels: E, C, & N.
- Contains no encyclopedic knowledge, no named entities, relatively small vocabulary, less multiword expressions and no lengthy sentences (9 words per sentence).
- Contradictions (86%) rely too much on negative words and antonyms [Lai and Hockenmaier, 2014].
- A benchmark for the SemEval-14 RTE task [Marelli et al., 2014a]: Trial (5%), Train (45%), and test (50%).
- 84% of crowd workers’ labels match the majority, i.e., gold labels.
SICK construction

<table>
<thead>
<tr>
<th>Original pair</th>
<th>Normalized pair</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S0a:</strong> A sea turtle is hunting for fish</td>
<td><strong>S0b:</strong> The turtle followed the fish</td>
</tr>
<tr>
<td><strong>S1a:</strong> A sea turtle is hunting for fish</td>
<td><strong>S1b:</strong> The turtle is following the fish</td>
</tr>
<tr>
<td><strong>S2a:</strong> A sea turtle is hunting for food</td>
<td><strong>S2b:</strong> The turtle is following the red fish</td>
</tr>
<tr>
<td><strong>S3a:</strong> A sea turtle is not hunting for fish</td>
<td><strong>S3b:</strong> The turtle isn’t following the fish</td>
</tr>
<tr>
<td><strong>S4a:</strong> A fish is hunting for a turtle in the sea</td>
<td><strong>S4b:</strong> The fish is following the turtle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normalized sentence pairs</th>
<th>Score</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1a: A sea turtle is hunting for fish</td>
<td>S2a: A sea turtle is hunting for food</td>
<td>4.5</td>
</tr>
<tr>
<td>S3a: A sea turtle is not hunting for fish</td>
<td>S1a: A sea turtle is hunting for fish</td>
<td>3.4</td>
</tr>
<tr>
<td>S4a: A fish is hunting for a turtle in the sea</td>
<td>S1a: A sea turtle is hunting for fish</td>
<td>3.9</td>
</tr>
<tr>
<td>S2b: The turtle is following the red fish</td>
<td>S1b: The turtle is following the fish</td>
<td>4.6</td>
</tr>
<tr>
<td>S1b: The turtle is following the fish</td>
<td>S3b: The turtle isn’t following the fish</td>
<td>4</td>
</tr>
<tr>
<td>S1b: The turtle is following the fish</td>
<td>S4b: The fish is following the turtle</td>
<td>3.8</td>
</tr>
<tr>
<td>S1a: A sea turtle is hunting for fish</td>
<td>S2b: The turtle is following the red fish</td>
<td>4</td>
</tr>
<tr>
<td>S1a: A sea turtle is hunting for fish</td>
<td>S3b: The turtle isn’t following the fish</td>
<td>3.2</td>
</tr>
<tr>
<td>S4b: The fish is following the turtle</td>
<td>S1a: A sea turtle is hunting for fish</td>
<td>3.2</td>
</tr>
<tr>
<td>S1b: The turtle is following the fish</td>
<td>S2a: A sea turtle is hunting for food</td>
<td>3.9</td>
</tr>
<tr>
<td>S1b: The turtle is following the fish</td>
<td>S3a: A sea turtle is not hunting for fish</td>
<td>3.4</td>
</tr>
<tr>
<td>S4a: A fish is hunting for a turtle in the sea</td>
<td>S1b: The turtle is following the fish</td>
<td>3.5</td>
</tr>
<tr>
<td>S1a: A sea turtle is hunting for fish</td>
<td>S1b: The turtle is following the fish</td>
<td>3.8</td>
</tr>
</tbody>
</table>
### SICK examples and stats

**SICK-1241**  
GOLD: neutral  
A woman is dancing and singing with other women  
A woman is dancing and singing in the rain

**SICK-341**  
GOLD: contradiction  
There is no girl with a black bag on a crowded train  
A girl with a black bag is on a crowded train

**SICK-8381**  
GOLD: entailment  
The young girl in blue is having fun on a slide  
The young girl in blue is enjoying a slide

<table>
<thead>
<tr>
<th>Relatedness</th>
<th>neutral</th>
<th>contradiction</th>
<th>entailment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,2) range</td>
<td>10%</td>
<td>0%</td>
<td>0%</td>
<td>10% (923)</td>
</tr>
<tr>
<td>[2,3) range</td>
<td>13%</td>
<td>1%</td>
<td>0%</td>
<td>14% (1373)</td>
</tr>
<tr>
<td>[3,4) range</td>
<td>28%</td>
<td>10%</td>
<td>1%</td>
<td>29% (3872)</td>
</tr>
<tr>
<td>[4,5] range</td>
<td>7%</td>
<td>3%</td>
<td>27%</td>
<td>37% (3672)</td>
</tr>
<tr>
<td>Total</td>
<td>56.86%</td>
<td>14.47%</td>
<td>28.67%</td>
<td>9840</td>
</tr>
</tbody>
</table>
The FraCaS dataset

The FraCaS test suite [Cooper et al., 1996] was an early attempt to creating a semantic benchmark for NLP systems.

- Contains 346 problems, 45% of which are multi-premised.
- Covers GQs, plurals, anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs and attitudes.
- Three-way annotated by the authors of the dataset.
- Contains some ambiguous sentences and a few erroneous problems.
- Requires almost no lexical or world knowledge.

Later, the FraCaS question-answer pairs where converted into an NLI format [MacCartney and Manning, 2007].
FraCaS NLI problems

- **FraCaS-26**  GOLD: entailment
  Most Europeans are resident in Europe
  All Europeans are people
  All people who are resident in Europe can travel freely within Europe
  Most Europeans can travel freely within Europe

- **FraCaS-61**  GOLD: undefined
  Both female commissioners used to be in business.
  Both commissioners used to be in business.

- **FraCaS-171**  GOLD: entailment
  John wants to know how many men work part time.
  And women.
  John wants to know how many women work part time.

- **FraCaS-87**  GOLD: entailment
  Every representative and client was at the meeting.
  Every representative was at the meeting.
Learning phase

The prover LangPro is (semi-automatically) trained on the NLI datasets [Abzianidze, 2016a].

**Adaptation:**

*NLI problems* \[\xrightarrow{\text{Prove}}\] \[\xrightarrow{\text{Adapt manually}}\] CCG parser + LLFgen + SG, KB, IR, PE

Used datasets: SICK-trial and FraCaS

**Development:**

Finding optimal values for certain parameters of the prover based on its performance on SICK-train.

**NB:** Only C&C parser is used in the learning phase in order to test LangPro for an unseen parser, EasyCCG, later.
Adaptation: negative cases

We avoid fitting to the data and adopting unsound and non-general solutions.

The problems that were not solved during the adaptation:

- Sentence is not recognised as of category $S$ or failed to be parsed

- The error is analysis is too specific to fix:
  
<table>
<thead>
<tr>
<th>At</th>
<th>most</th>
<th>ten</th>
<th>commissioners</th>
<th>spend</th>
<th>time</th>
<th>at</th>
<th>home</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(S/S)/NP$</td>
<td>$N/N$</td>
<td>$N/N$</td>
<td>$N$</td>
<td>$(VP/PP)/NP$</td>
<td>$N$</td>
<td>$PP/NP$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- Lexical relation is context dependent:
  
  SICK-4505 GOLD: entailment
  The doctors are healing a man
  The doctor is helping the patient

  SICK-384 GOLD: entailment
  A white and tan dog is running through the tall and green grass
  A white and tan dog is running through a field
Adaptation: positive cases

The problems that were solved by upgrading one of the components of the prover:

- Treat **few** as ↓ in its 1st arg (**absolute** reading):
  - FraCaS-76 GOLD: entailment
    - Few committee members are from southern Europe
    - Few female committee members are from southern Europe

- Introduce **fit** ⊑ **apply** and **food** ⊑ **meal**:
  - SICK-4734 GOLD: entailment
    - A man is **fitting** a silencer to a pistol
    - A man is **applying** a silencer to a gun
  - SICK-5110 GOLD: entailment
    - A chef is preparing some **food**
    - A chef is preparing a **meal**
Development phase

Optimal values of the following parameters are searched:

- The number of word senses to consider at the same time;
- The upper bound for the number of rule applications;
- Whether to use a term aligner:
  - **Weak aligner** aligns everything but terms of type np:
    - SICK-1022  
    - **GOLD**: contradiction  
    - A woman is **wearing sunglasses of large size and is holding newspapers in both hands**  
    - There is no woman **wearing sunglasses of large size and holding newspapers in both hands**
    - SICK-727  
    - **GOLD**: contradiction  
    - The man in a grey t-shirt is **sitting on a rock in front of the waterfall**  
    - There is no man **in a grey t-shirt sitting on a rock in front of the waterfall**
  - **Strong aligner** aligns everything but terms of type terms of type np with ↓arg.
    - SICK-423  
    - **GOLD**: contradiction  
    - **Two men are not holding fishing poles**  
    - Two men are holding fishing poles
  - Efficiency criterion of tableau rules.
Efficiency criterion

Tableau rules have the following properties:

- Non-branching or branching (so called, $\alpha$ or $\beta$ rules);
- Semantic equivalence vs proper entailment;
- Consuming (so called, $\gamma$ rule) vs non-consuming;
- Producing (so called, $\delta$ rule) vs non-producing.

An example of an efficiency criterion:

$$EC = \langle \text{nonBr}, \text{semEqui}, \text{nonConsum}, \text{nonProd} \rangle$$

An efficiency vectors based on the $EC$ efficiency criterion:

$$V_{EC}(\land_T) = 1111$$
$$V_{EC}(\lor_T) = 0111$$
$$V_{EC}(\exists_T) = 1110$$
$$V_{EC}(\exists_F) = 0001$$

What is the optimal efficiency criterion?
Greedy search for optimal parameters

<table>
<thead>
<tr>
<th>Acc%</th>
<th>Prec%</th>
<th>Rec%</th>
<th>Sense</th>
<th>Efficiency criterion</th>
<th>Aligner</th>
<th>RAL</th>
<th>Parser</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.09</td>
<td>98.5</td>
<td>43.6</td>
<td>1</td>
<td>[nonP, nonB, equi, nonC]</td>
<td>No</td>
<td>200</td>
<td>C&amp;C</td>
</tr>
<tr>
<td>76.42</td>
<td>98.3</td>
<td>46.8</td>
<td>1-5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>76.89</td>
<td>97.8</td>
<td>48.1</td>
<td>All</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>78.44</td>
<td>97.9</td>
<td>51.7</td>
<td>-</td>
<td>[equi, nonB, nonP, nonC]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>79.33</td>
<td>97.9</td>
<td>53.8</td>
<td>-</td>
<td>-</td>
<td>Weak</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>81.5</td>
<td>97.7</td>
<td>59.0</td>
<td>-</td>
<td>-</td>
<td>Strong</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>81.53</td>
<td>97.7</td>
<td>59.1</td>
<td>-</td>
<td>-</td>
<td>Strong</td>
<td>400</td>
<td>-</td>
</tr>
<tr>
<td>81.38</td>
<td>98.0</td>
<td>58.5</td>
<td>-</td>
<td>-</td>
<td>Strong</td>
<td>400</td>
<td>EasyCCG</td>
</tr>
<tr>
<td><strong>82.6</strong></td>
<td><strong>97.7</strong></td>
<td><strong>61.6</strong></td>
<td>-</td>
<td>-</td>
<td>Strong</td>
<td>400</td>
<td><strong>Both</strong></td>
</tr>
</tbody>
</table>

The results are given on the SICK-train problems.

**FraCaS-21**

GOLD: entailment

The residents of member states have the right to live in Europe
All residents of member states are individuals
Every individual who has the right to live in Europe can travel freely within Europe
The residents of member states can travel freely within Europe
The results are given on the SICK-train problems.
Solving FraCaS [Abzianidze, 2016b]

<table>
<thead>
<tr>
<th>Gold\ccLP</th>
<th>yes</th>
<th>no</th>
<th>unk</th>
<th>P = .97, R = .71, Acc = .81</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>51</td>
<td>0</td>
<td>19 + 4</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>unk</td>
<td>1</td>
<td>0</td>
<td>44 + 6</td>
<td></td>
</tr>
</tbody>
</table>

LangPro with C&C + LangPro with EasyCCG = LangPro

<table>
<thead>
<tr>
<th>Gold\easyLP</th>
<th>yes</th>
<th>no</th>
<th>unk</th>
<th>P = .96, R = .70, Acc = .80</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>52</td>
<td>0</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>unk</td>
<td>2</td>
<td>0</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gold\LP</th>
<th>yes</th>
<th>no</th>
<th>unk</th>
<th>P = .96, R = .81, Acc = .87</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>60</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>unk</td>
<td>2</td>
<td>0</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>

FraCaS-109  
**GOLD:** contradiction  
**LP:** entailment

**Just one** accountant attended the meeting

**Some accountants** attended the meeting
Related work (FraCaS)

[MacCartney and Manning, 2008] and [Angeli and Manning, 2014] employ a natural logic that is driven by sentence edits.


[Mineshima et al., 2015] also uses the Boxer-style translation but some HOGQs are treated as higher-order terms. Their inference system is implemented in the proof assistant Coq.

[Tian et al., 2014] and [Dong et al., 2014] uses abstract denotations obtained from DCS trees [Liang et al., 2011]:

\[ \text{man} \subseteq \pi_{\text{subj}}(\text{read} \cap (W_{\text{subj}} \times \text{book}_{\text{obj}})) \]

Comparison on FraCaS

<table>
<thead>
<tr>
<th>Sec (Sing/All)</th>
<th>Single-premised (Acc %)</th>
<th>Overall (Acc %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BL</td>
<td>NL07,08</td>
</tr>
<tr>
<td>1 GQs (44/74)</td>
<td>45</td>
<td>84</td>
</tr>
<tr>
<td>2 Plur (24/33)</td>
<td>58</td>
<td>42</td>
</tr>
<tr>
<td>5 Adj (15/22)</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>9 Att (9/13)</td>
<td>67</td>
<td>56</td>
</tr>
<tr>
<td>1,2,5,9 (92/142)</td>
<td>50</td>
<td>-</td>
</tr>
</tbody>
</table>

**NL07** [MacCartney and Manning, 2007], **NL08** [MacCartney and Manning, 2008], **NLI** [Angeli and Manning, 2014], **LS** [Lewis and Steedman, 2013], **M15** [Mineshima et al., 2015], **T14a** [Tian et al., 2014] and **T14b** [Dong et al., 2014]

Advantages of our approach over the related ones include:

- Reasoning (with the semantic tableau) over multiple-premises;
- Logical forms close to surface forms;
- Underlying expressive high-order logic.
Curing SICK [Abzianidze, 2015]

Mainly the usage of WordNet and noisy gold labels are blamed for false proofs.

<table>
<thead>
<tr>
<th>ID</th>
<th>G/LP</th>
<th>Premise</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1405</td>
<td>N/E</td>
<td>A <strong>prawn</strong> is being cut by a woman</td>
<td>A woman is cutting <strong>shrimps</strong></td>
</tr>
<tr>
<td>4443</td>
<td>N/E</td>
<td>A man is singing to a <strong>girl</strong></td>
<td>A man is singing to a <strong>woman</strong></td>
</tr>
<tr>
<td>2870</td>
<td>N/C</td>
<td>Two people are riding a <strong>motorcycle</strong></td>
<td>Nobody is riding a <strong>bike</strong></td>
</tr>
<tr>
<td>8913</td>
<td>N/C</td>
<td>A couple is not looking at a map</td>
<td>A couple is looking at a map</td>
</tr>
<tr>
<td>363</td>
<td>C/C</td>
<td>P: A soccer ball is not rolling into a goal net</td>
<td>C: A soccer ball is rolling into a goal net</td>
</tr>
</tbody>
</table>
False neutrals

Reason for false neutrals are knowledge sparsity (ca 50%), a lack of rules (ca 25%), wrong labels and parsing mistakes.

<table>
<thead>
<tr>
<th>ID</th>
<th>G/LP</th>
<th>Premise</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>4974</td>
<td>E/N</td>
<td>Someone is holding a <strong>hedgehog</strong></td>
<td>Someone is holding a <strong>small animal</strong></td>
</tr>
<tr>
<td>6258</td>
<td>E/N</td>
<td>P: A <strong>policeman</strong> is sitting on a <strong>motorcycle</strong></td>
<td>C: The cop is sitting on a <strong>police bike</strong></td>
</tr>
<tr>
<td>4553</td>
<td>E/N</td>
<td>P: A man is emptying a <strong>container made of plastic</strong></td>
<td>C: A man is emptying a <strong>plastic container</strong></td>
</tr>
<tr>
<td>4720</td>
<td>E/N</td>
<td>A <strong>monkey</strong> is practicing martial arts</td>
<td>A <strong>chimp</strong> is practicing martial arts</td>
</tr>
<tr>
<td>6447</td>
<td>C/N</td>
<td>P: [A small boy [in a yellow shirt]] is laughing on the beach</td>
<td>C: There is no small boy [in a yellow shirt [laughing on the beach]]</td>
</tr>
</tbody>
</table>
## Comparison on SICK

<table>
<thead>
<tr>
<th>SemEval-14 systems</th>
<th>Prec%</th>
<th>Rec%</th>
<th>Acc%</th>
<th>(+LP)</th>
<th>NWS%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (majority)</td>
<td>-</td>
<td>-</td>
<td>56.69</td>
<td>(+0.65)</td>
<td>39.7</td>
</tr>
<tr>
<td>Illinois-LH</td>
<td>81.56</td>
<td>81.87</td>
<td>84.57</td>
<td>(+0.65)</td>
<td>72.8</td>
</tr>
<tr>
<td>ECNU</td>
<td>84.37</td>
<td>74.37</td>
<td>83.64</td>
<td>(+1.77)</td>
<td>72.7</td>
</tr>
<tr>
<td>UNAL-NLP</td>
<td>81.99</td>
<td>76.80</td>
<td>83.05</td>
<td>(+1.48)</td>
<td>71.2</td>
</tr>
<tr>
<td>SemantiKLUE</td>
<td>85.40</td>
<td>69.63</td>
<td>82.32</td>
<td>(+2.84)</td>
<td>71.5</td>
</tr>
<tr>
<td>The Meaning Factory</td>
<td>93.63</td>
<td>60.64</td>
<td>81.59</td>
<td>(+2.78)</td>
<td>73.0</td>
</tr>
<tr>
<td>UTexas (Prob-FOL)</td>
<td><strong>97.87</strong></td>
<td>38.71</td>
<td>73.23</td>
<td>(+9.44)</td>
<td>62.5</td>
</tr>
<tr>
<td><strong>LangPro</strong></td>
<td><strong>97.35</strong></td>
<td>60.31</td>
<td>82.14</td>
<td></td>
<td>74.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RTE systems</th>
<th>Acc%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob-FOL</td>
<td>76.52</td>
</tr>
<tr>
<td>Prob-FOL* +Rules</td>
<td><strong>85.10</strong></td>
</tr>
<tr>
<td>Nutcracker+PPDB</td>
<td>79.60</td>
</tr>
<tr>
<td>ABCNN-3</td>
<td><strong>86.20</strong></td>
</tr>
<tr>
<td>LSTM RNN+SNLI</td>
<td>80.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gold\System</th>
<th>E</th>
<th>C</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entailment</td>
<td>2</td>
<td>–2</td>
<td>0</td>
</tr>
<tr>
<td>Contradiction</td>
<td>–2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>–1</td>
<td>–1</td>
<td>1</td>
</tr>
</tbody>
</table>
“Hard” problems

The problems from SICK-test that were proved correctly by both ccLangPro and easyLangPro but failed by all the top five systems at the SemEval-14 task.

<table>
<thead>
<tr>
<th>ID</th>
<th>G</th>
<th>Text</th>
<th>Hypothesis</th>
</tr>
</thead>
</table>
| 247 C | T: The woman is not wearing glasses or a headdress  
   H: A woman is wearing an Egyptian headdress |
| 406 E | T: A group of scouts are hiking through the grass  
   H: People are walking |
| 2895 C | The man isn’t lifting weights  
   The man is lifting barbells |
| 3527 E | T: A person is jotting something with a pencil  
   H: A person is writing |
| 3570 C | The piece of paper is not being cut  
   Paper is being cut with scissors |
| 3608 N | T: A monkey is riding a bike  
   H: A bike is being ridden over a monkey |
| 3806 E | A man in a hat is playing a harp  
   A man is playing an instrument |
| 4479 E | The boy is playing the piano  
   The boy is playing a musical instrument |
Introducing a new tableau rule

Let us add a new rule to Natural Tableau and LangPro:

We want introduce a rule in order to account for the entailment:

**GOLD:** entailment
Most women are working
Most women are rich
There is a woman who is working and is rich

This rule will help:

\[
\begin{array}{c}
\text{MOST2} \\
[\overrightarrow{M}_1]: \text{most } N_n \ W_{vp}: [] : \top \\
[\overrightarrow{M}_2]: \text{most } N_n \ R_{vp}: [] : \top \\
\vdash N_n: [c_e]: \top \\
[\overrightarrow{M}_1]: W_{vp}: [c_e]: \top \\
[\overrightarrow{M}_2]: R_{vp}: [c_e]: \top \\
\hline
\text{c}_e \text{ is fresh and } W \neq R
\end{array}
\]
Conclusion

Natural Tableau is a wide-coverage but still logic-based reasoning system inspired by Natural Logic.

It represents a proof-theoretic approach to NLI.

Natural tableau was successfully scaled up for the NLI task:

CCG parser + LLFgen + theorem prover

Pros and cons of Natural Tableau:

✓ Employs higher-order logic to model linguistic semantics;
✓ Allows deep logical and shallow (e.g. monotonicity) reasoning;
✓ Getting logical form is similar to syntactic parsing;
✗ Heavily hinges on CCG parsing;
✓ Proofs are highly reliable (≤3% false proofs);
✗ Suffers from multi-sense words;
✗ No fully automated learning from data yet;
✓ Its decision procedure is transparent and explanatory;
Future work

There are really many directions for future work:

- Explore different types of RTE data, e.g., the newswire or human generated data [Bowman et al., 2015];
- Incorporate more knowledge in KB, e.g., paraphrase database [Ganitkevitch et al., 2013].
- Model different phenomena: comparatives, anaphora, cardinals, etc.
- Pairing with distributional semantics: \( R(w_1, w_2, r) \) and weighted closure branches;
- Acquisition of lexical knowledge: abductive reasoning;
- Generate LLFs from Universal Dependency trees
  + the Universal Semantic Tagging [?]
  \( \rightarrow \) Multilingual Natural Tableau
Inference to the best explanation

1. \textbf{person}_n : [p_e] : \top

2. \textbf{hedgehog}_n : [a_e] : \top

3. small_{n,n} \textbf{animal}_n : [a_e] : \perp

4. \textbf{hedgehog}_n : [h_e] : \top

5. \textbf{hold}_{np,vp} : [h_e, p_e] : \top

6. \textbf{man}_n : [m_e] : \top

7. \textbf{box}_n : [b_e] : \top

8. \textbf{chicken}_n : [c_e] : \top

4. \textbf{into}_b_e : \textbf{put}_{np,pp,vp} : [c_e, m_e] : \top

5. \textbf{food}_n : [f_e] : \top

6. \textbf{from}_b_e : \textbf{remove}_{np,pp,vp} : [f_e, m_e] : \top
Thank you for coming here in the early mornings and listening me repeating tableau, tableau, tableau, . . . , tableau!


References III


