Banks Meeting: PropBanks

2009-02-24

Martha Palmer

http://www.bbn.com/NLP/OntoNotes
PropBank - *press*

- **press** Roleset press.01 "apply pressure":
  
  Verbnet: push 12.1, force 59.1
  
  - Roles:
    - Arg0: presser
    - Arg1: pressed on
    - Arg2: goal, pressed for
  
    - [ArgM-DIS But] [Arg0 Ford, Jaguar's biggest shareholder since lifting its stake to 10.4% this week], is **pressing** [ArgM-ADV harder] [Arg2 for talks with Sir John.]

- **press** Roleset press.02 "move":
  
  - Roles:
    - Arg0:mover
    - Arg1: moved
    - ArgM-DIR: ahead, usually

- WordNet – 14 verb senses
Functional tags for modifiers or adjuncts

- Variety of ArgM’s:
  - **TMP** - when? *yesterday, 5pm on Saturday, recently*
  - **LOC** - where? *in the living room, on the newspaper*
  - **DIR** - where to/from? *down, from Antarctica*
  - **MNR** - how? *quickly, with much enthusiasm*
  - **PRP/CAU** - why? *because …, so that …*
  - **REC** - himself, themselves, each other
  - **ADV** - hodge-podge, miscellaneous, “nothing-fits!”
  - **PRD** - this argument refers to or modifies another
Adding Lemmas

- New corpora often require Frame Files for previously unseen lemmas or edits to existing Frame Files
- Adding Frame Files – Currently 4,184 (5600+ senses)

<table>
<thead>
<tr>
<th>Data</th>
<th>New Frame Files</th>
<th>New Frame Sets</th>
<th>Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Broadcast News</td>
<td>80</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>E Broadcast Conv</td>
<td>86</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>E/C Parallel TB</td>
<td>87</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>From VerbNet</td>
<td>79</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Added</strong></td>
<td><strong>344</strong></td>
<td><strong>70</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>
English PropBank YR 3 Status

- Corpora, Type coverage, 100%
  - Lemmas: 4,184 (up from 3100 WSJ-Newswire)
  - Framesets: 5600+ (multiple subcategorization frames for a lemma)

- Corpora, token coverage, 100%*
  - Y1 (NW, 300K) delivered
  - Y2 (BN, 200K) delivered – 33,800 instances completed
  - YR2 & YR3 (ECTB, 250K) – 30,000 instances completed
  - Y3 (BC, 200K) delivered – 27,600 instances, completed
    - Went quite a bit more slowly
    - Will also PropBank P2.5 (parallel MT evaluation data)
  - Y4 (NG/WL) Data being selected, with a goal of improving coverage for targeted lexical items
    - *does not yet include YR 2/3 eventive nouns/nominalizations
## CPB: Milestones

<table>
<thead>
<tr>
<th>Version</th>
<th>CPB1.0</th>
<th>CPB2.0</th>
<th>CPB.xx</th>
<th>CPB.xx</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>2005</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td><strong>Words</strong></td>
<td>250K</td>
<td>500K</td>
<td>800K</td>
<td>950K</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Xinhua</td>
<td>+Sinorama</td>
<td>+BN (300k)</td>
<td>+ BC (150K)</td>
</tr>
<tr>
<td><strong>OntoNotes Data</strong></td>
<td>partial</td>
<td>partial</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Predicates</strong></td>
<td>4,865</td>
<td>11,765</td>
<td>14,128</td>
<td>15,192</td>
</tr>
<tr>
<td><strong>Framesets</strong></td>
<td>5,298</td>
<td>12,555</td>
<td>15,077</td>
<td>??</td>
</tr>
<tr>
<td><strong>Frameset/pred</strong></td>
<td>1.09</td>
<td>1.06</td>
<td>1.07</td>
<td>??</td>
</tr>
<tr>
<td><strong>Propositions</strong></td>
<td>37,183</td>
<td>92,180</td>
<td>143,303</td>
<td>174,144</td>
</tr>
</tbody>
</table>
Using Parallel Propbanks to enhance word-alignments – Choi, Palmer, Xue

- **GIZA++ Alignments of Xinhua and Sinorama**

- Use predicate-argument structures to verify and expand word-alignments for verb predicates
Results

- Two approaches
  - Top-down – start with GIZA++ matched predicates, *verify* by checking argument matches, use thresholds
  - Bottom-up – start with matching arguments, given enough matching arguments assume predicates match - *expansion*

- Test with NIST MT eval data (thanks to ISI)
  - 273 verb types, 100 sentences

- Top-down approach effective with Sinorama
  - GIZA++ generates 129 mappings, but w/ acc 57.76
  - T-D filters out 67 for 69 total, but acc goes up to 78.09

- Bottom-up approach improves Xinhua
  - Add 22 new mappings w/ acc 50.38
  - Thresholds filter out 4, acc goes up to 63.89