THE GRONINGEN MEANING BANK

BUILDING A LARGE SEMANTICALLY ANNOTATED CORPUS

Johan Bos
21/11/2013
JSSP2013
Povo, Trento
The departure of Mr Hogan, who originally moved to British Midland as service director from Hertz International in 1997, surprised aviation analysts, as it was believed that he had been brought into the senior executive team of the airline, as part of the group's management succession planning.

He played a leading role in the strategic planning for the rebranding of the airline as BMI in preparation for its entry this year into the scheduled long haul market with the launch of services from Manchester to the US.

BMI has taken on the costs of entry into the North Atlantic market at an unfortunate time, as airlines in North America are facing the toughest conditions for 20 years with many carriers plunging into loss.

BMI, in which Lufthansa of Germany and SAS Scandinavian Airlines each own stakes of 20 per cent, suffered a 25 per cent fall in pre-tax profits last year from £11.1m ($15.7m) to £8.2m on a turnover that grew 16.5 per cent to £739.2m.

In the first six months this year it is understood that passenger volumes have fallen by around two per cent. The share of available seats filled, the load factor, has declined by around two percentage points, but this has been offset by a strong increase in yields, or average fare levels, by more than ten per cent.

BMI's move to tighten its management structure follows a warning on Monday from British Airways that the industry faces challenging months with a "more difficult" winter ahead.
SYNTACTIC ANNOTATION

Examples:

- Penn TB
- Alpino
- Lingo Redwoods
- Negra
- Praque DT

and many more!
SEMANTIC ANNOTATION

Examples:
The founder of Pakistan's Nuclear department has admitted he transferred Technology, Libya and North Korea.

Propositions:
- e1: admit
  - Arg0: Pakistan's nuclear department
  - Arg1: Abdul Qadeer Khan

- e2: founder
  - Arg0: Pakistan's nuclear department
  - Arg1: Abdul Qadeer Khan

- e3: Pakistan's nuclear department
- e4: Abdul Qadeer Khan
- e5: nuclear technology
- e6: transfer
  - Arg0: Pakistan's nuclear department
  - Arg1: Abdul Qadeer Khan
  - Arg2: to

Coreference:
- e2 = e4
- e2 = e7
Semantic parsing is now at a point where syntactic parsing was twenty years ago:

- rule-based approaches dominate
- no robust interpretation models exist
- semantic tasks often carried out in isolation
- lack of large annotated linguistic resources
OUR GOAL

• Syntactic parsing made a breakthrough by developing treebanks in the late 1980s
• We want to develop a large semantically annotated resource and cause a similar breakthrough for semantics!

we are ambitious!
OUR QUESTIONS

• What should such an annotated resource look like?
• How can we learn from earlier attempts?
• What semantic theory should we adopt?
• What phenomena should be covered?
• What size should it be?
• How do we build it?
THE GRONINGEN MEANING BANK

- large (English) corpus of public domain texts
- deep (logical), not shallow semantics
- automatically produced, manually corrected
- integrates various phenomena in one formalism
- linguistically motivated, theoretically solid (CCG/DRT)
INGREDIENTS OF THE MEANING BANK

- PTB part of speech tags
- Animacy classes
- Named entity types
- WordNet senses
- VerbNet roles
- CCG derivations (syntax)
- DRT boxes (semantics)
- SDRT rhetorical relations (discourse)
SEMANTIC PHENOMENA

- Collective/distributive NPs
- Scope (negation, quantifiers)
- Anaphoric pronouns
- Word senses (WordNet)
- Thematic Roles (VerbNet)
- Verb phrase ellipsis
- Presuppositions
- Tense and aspect
- neo-Davidsonian events
- Rhetorical relations
- Background knowledge
MEANING BANK DEVELOPMENT
1  We PRP VBP 're VBP (S[dc]\NP)/(S[adj]\NP) sorry RB [Patient] S[adj]\NP . (S[dc]\S[dc])
2  You PRP VBP have VBP (S[dc]\NP)/(S[pt]\NP) reached VBN [Theme,Agent] (S[pt]\NP)/NP an DT JJ NN number NN N . S[dc]\S[dc]
3  Please VB [Theme,Agent] (S[b]\NP)/(S[b]\NP) rotate VB [Theme,Agent] (S[b]\NP)/NP your PRPS NN phone NN 90 CD degrees NNS and CC conj try VB [Experience] S[b]\NP again RB (S[b]\NP)/(S[b]\NP) .
THE COMPOSITIONAL SEMANTICS LAYER
THE COMPOSITIONAL SEMANTICS LAYER

250
CD
NP/N
λν0. λν1. (x3 ; (v0 @ x3); (v1 @ x3))
| x3 | l|x3| = 250 |

film-coated
JJ
N/N
λν0. λν1. (film-coated(v1) ; (v0 @ v1))

| film-coated(v1) |

| tablet(v0) |

| tablets |

| N |

| NNS |

| tablet(v0) |

250 film-coated tablets
NP
λν0. (x2 ; (v0 @ x2))
| x2 | l|x2| = 250 |

| film-coated(x2) |

| tablet(x2) |
A neo-Davidsonian style sentence DRS, integrating semantic information

Fertility was not reduced
S[dcl]
λv0.

negation
WordNet
VerbNet
## GMB: CORPUS SIZE

<table>
<thead>
<tr>
<th>genre</th>
<th>texts</th>
<th>sentences</th>
<th>words</th>
<th>s/t</th>
<th>w/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice of America</td>
<td>9,123</td>
<td>56,533</td>
<td>1,224,262</td>
<td>6.2</td>
<td>21.7</td>
</tr>
<tr>
<td>CIA world factbook</td>
<td>514</td>
<td>4,435</td>
<td>112,519</td>
<td>8.6</td>
<td>25.3</td>
</tr>
<tr>
<td>Aesop’s Fables</td>
<td>225</td>
<td>959</td>
<td>23,223</td>
<td>4.3</td>
<td>24.2</td>
</tr>
<tr>
<td>jokes</td>
<td>123</td>
<td>448</td>
<td>7,638</td>
<td>3.6</td>
<td>17.0</td>
</tr>
<tr>
<td>MASC</td>
<td>35</td>
<td>292</td>
<td>6,987</td>
<td>8.3</td>
<td>32.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10,020</strong></td>
<td><strong>62,667</strong></td>
<td><strong>1,374,629</strong></td>
<td><strong>6.3</strong></td>
<td><strong>21.9</strong></td>
</tr>
</tbody>
</table>
- Annotations are produced by machines, corrected by humans
- A lot of it is done by NLP tools: C&C tools, Boxer
- But substantial **human** annotation cannot be avoided
TRADITIONAL ANNOTATION
CROWDSOURCING

- Outsourcing tasks to a distributed group of people
- The internet provides infrastructure
- Two crowds:
  - Expert Annotators
  - GWAP players
GAMES WITH A PURPOSE

Play the Games, Change the Web.

With Duolingo you learn a language for free while helping to translate the web.
Enjoy a language game with Wordrobe!
There are many games to choose from!

Play Twins

the other games

Fascinated by language? Play!

Synchronize wordrobe with your social network
Facebook sync  Twitter sync

Play Wordrobe

Top scores

1  Aristotle | 4526 points
2  Kilian   | 2572 points
3  Nynke   | 1542 points
4  Valerio  | 1083 points
5  Leo van Maanen | 1024 points
6  Potato   | 631 points
7  Noortje  | 581 points
8  MichaelHahn | 503 points

Choose another game

WWW.WORDROBE.ORG
not a single game, but a series of games that share same structure and scoring strategies

- each linguistic phenomenon that requires annotation corresponds to a different game
- every game consists of multiple-choice questions
- each question is presented by a text fragment plus a (small) number of possible answers
- these questions (and answers) are automatically generated from the corpus
A SCORE WITH A TWIST

- In addition, Wordrobe players can take risks and bet on the correctness of an answer.
- The higher the bet, the more points you can win (or lose).

Senses

One Afghan soldier was killed and four others, **including** a U.S. soldier, were injured in the fighting.

- have as a part, be made up out of
- consider as part of something
- add as part of something else – put in as part of a set, group, or category
- allow participation in or the right to be part of – permit to exercise the rights, functions, and responsibilities of (synonyms: admit, let in)

Place your bet: low [ ] high [ ]

[answer] [skip]
Every answer increases the score of a player.

The more overlap of a player’s answer with other players, the higher the score.

Total scores for a game are calculated over answers given in the last N days (currently N=50).
WHY WILL PEOPLE PLAY WORDROBE?

- unlocking achievements
- outperforming other players
- learning about language
- social contacts
- killing time
- help science
GMB IN NUMBERS

- 1,285 wordrobe players,
- 62,102 answers (6,982 “bits of wisdom”)
- 15 expert annotators
- 30,103 “bits of wisdom”
- 4 stable releases
- 164 downloads
1. Data-driven formal semantics
2. Multi-lingual meaning banking
3. Generating from formal meanings
4. Learning semantic parsing
<table>
<thead>
<tr>
<th>600</th>
<th>mg</th>
<th>telbivudine</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>600</td>
<td>mg</td>
<td>telbivudine</td>
</tr>
<tr>
<td>CD</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>NP/N</td>
<td>N/N</td>
<td>N</td>
</tr>
</tbody>
</table>
LEARNING SEMANTIC PARSING

\[ \lambda v_0. \lambda v_1. ( \begin{array}{c} x_3 \\ l|x_3| = 30 \end{array} ; (v_0 \@ x_3) ; (v_1 \@ x_3) ) \]

\[ \lambda v_0. \lambda v_1. ( \begin{array}{c} x_3 \\ mg(x_3) \\ of(v_1, x_3) \end{array} ; (v_0 \@ v_1) ) \]

\[ \lambda v_0. \begin{array}{c} \text{telbivudine} \\ \text{telbivudine}(v_0) \end{array} \]
Current results, algorithm of Zettlemoyer & Collins (2007)
7-word sentences: 114/193 semantic parses correct (0.59)
SPECTACULARLY WRONG OUTPUT

**Gold Standard**

Hycamtin helps to destroy tumours.

\[ S[\text{dcl}] \]
\[ \lambda v. (x_2 \ p_4 \ e_5 \ ; (v_0 @ e_5) ) \]

- named(x_2, hycamtin, nam)
- p_4: x_7 \ e_8
- tumour(x_7)
- destroy(e_8)
- Agent(e_8, x_2)
- Patient(e_8, x_7)
- help(e_5)
- Agent(e_5, x_2)
- Theme(e_5, p_4)

**Produced by crossboxer**

“Hycamtin helps to destroy itself on tumours”
The GMB is a public domain corpus, free for research
Formal semantic annotation (not semi-formal)
Units of annotations are texts (not sentences)
Anyone can contribute! (experts and street folk)

http://gmb.let.rug.nl