Learning relational meanings from situated caregiver-child interaction
A computational approach

Barend Beekhuizen\textsuperscript{1}, Afsaneh Fazly\textsuperscript{2}, Aida Nematzadeh\textsuperscript{2} & Suzanne Stevenston\textsuperscript{2}

\textsuperscript{1}Leiden University \quad \textsuperscript{2}University of Toronto

13 November 2012
1 The problem of learning relational meaning
   - Learning word meaning
   - Relational meaning
   - Our approach

2 A computational approach
   - Fazly, Alishahi & Stevenson (2010)
   - Data

3 Bootstraps and biases
   - Addressing the ‘missing meaning’ problem
   - Addressing the ‘missing words’ problem
   - Addressing the ‘too much meaning’ problem
   - Other possible bootstraps
   - An overview: what helps, what doesn’t
   - Whither?
Suppose you can segment utterances into words
& suppose you understand that others have communicative intentions when they use language
& suppose you can partially understand these communicative intentions without understanding language
... can you learn the mappings between words and the objects and situations they refer to?
People typically think of object-referential meaning when they talk about meaning (‘ball’ and ‘dog’ and ‘chair’)

What about reference to relations between objects (relational meanings)?

E.g., ‘being-underneath’; ‘exerting force upon’; ‘moving w.r.t.’; ‘having a similar shape’

Verbs, prepositions, relational nouns,
Why is it hard?

- Words with relational meaning (RM) are thought to be hard to learn.
- Why?¹
  - Stability of RM: not stable over time (as opposed to objects)
  - Quantity of RM and hypothesis space: many relations holding between objects (objects: more limited)
  - Perceptibility of RM (beliefs, attitudes, perception)

---

Proposed solutions

- Syntactic bootstrapping\(^2\)
- Constraints (in particular: mutual exclusivity)\(^3\)
- Socio-pragmatic bootstrapping\(^4\)
- All of the above \(^5\)

\(^2\)Gleitman (1990)

\(^3\)Markman, E. M. (1994). ‘Constraints on word meaning in early language acquisition’. Lingua, 92, 199227.


Revisiting the claims

• Before trying to **solve** the problem: estimate its magnitude (hasn’t been done since Gleitman (1990), though there is a rising interest in observational data).  

• Get a more detailed picture of the problems.

• Then: using computational modeling techniques to test some of the hypothesized solutions.

---


The issue of the data

- Lots of work in experimental settings: problems with ecological validity
- Esp.: underestimation of hypothesis space, noisiness (Medina et al. 2011)
- So: we look at observational data of less constrained caregiver-child interaction
The pegs and holes game

- Dyads of mothers and daughters (16 mo) playing games of putting pegs in holes
- Mothers instructing children verbally (in Dutch)
- 32 dyads, approximately 5 minutes each: 157 minutes in total

Figure: The toy and the twelve blocks
The pegs and holes game

- Getting information from the video’s:
  - Two coders coded behavior, spatial states and involved objects according to a coding schema
  - Within 3-second intervals
  - Format: predicate-argument structures, e.g., grab(mother, yellow-square-block)
  - With good inter- and intra-coder agreement (most $\kappa > 0.8$)
  - I transcribed all speech
The pegs and holes game

- Getting information from the video’s:
  - Two coders coded behavior, spatial states and involved objects according to a coding schema
  - Within 3-second intervals
  - Format: predicate-argument structures, e.g., grab(mother, yellow-square-block)
  - With good inter- and intra-coder agreement (most $\kappa > 0.8$)
  - I transcribed all speech

- Resulting corpus:
  - 157 minutes of behavior-coded and speech-transcribed material
  - 2492 utterances, 7842 word tokens (480 types, 355 lemmas)
  - 8464 behavioral predicates
  - Other information: fit of block and hole.
Words and things

- Expressing words in coded meaning:
  - *pakken* ‘to grab’ means *grab*
  - *stoppen* ‘to put (sth. into sth.)’ means \{move, in\}
  - *geel* ‘yellow’ means *yellow*
  - *op* ‘on’ means *on*

- Using these representations, we can check if the (correct) feature occurs in interval of utterance.

- Also: if word occurs when the feature is present

- Using these descriptive statistics, we encounter three problems for the learner.
Problem 1: missing meaning

- **Meaning is not present in situational context** of word:
  - **utterance**: *You go grab the block!*
    - **situation**: grabbing takes place 7 seconds later.
  - **utterance**: *Hey, don’t take the lid off!*
    - **situation**: child is pulling at the lid, but doesn’t succeed in taking it off

- Calculate percentage of utterances containing a word in which the correct feature is present.
- E.g.: *pakken* ‘to grab’: in 58% of cases is *grab* present
- Globally (proportions): words for colors/shapes (0.75) > verbs (0.59), object nouns (0.57) and spatial terms (0.53)
Problem 2: missing words

- **Word expressing** an aspect of the situational **context** is **not present** in the utterance:
  - **utterance**: *Good girl!*
    - **situation**: child puts block in bucket
  - **utterance**: *Now it’s off!*
    - **situation**: child grabs lid and moves it off of the bucket

- Calculate how often word is used when meaning is present.
- This problem is **huge**: meaning will be associated with lots of other words.
- Spatial states (0.06) > verbs (0.02) > objects (0.009), colors/shapes (0.008)
- (Problem seems bigger for non-relational meaning than for relational meaning)
Problem 3: too much meaning

- Irrelevant features co-occur often with word:
  - utterance: *That sure will fit there*
  - situation: child is fitting block in right hole, but other relations are there too: child positioning the block, block being near to the hole, child holding the block etc.

- Partially due to nature of the data: restricted nature of agents (child & mother), patients (blocks, lid) and locations (bucket, hole, floor).
Three problems

- Amount of **referential uncertainty**, **feature non-independence** and **noise** seems bigger than in lab settings and semantic datasets built from child-directed language (with **synthetic meaning**).
- So, how would a computational learner behave facing this data?
- What known mechanisms can help the learner reduce the problems
Starting point: model of Fazly, Alishahi & Stevenson (2010); FAS

Assume that the input consists of a situation $S$ and an utterance $U$

Let $S$ be a set of features $f_1 \ldots f_n$ present in the situational context

Let $U$ be a set of words $w_1 \ldots w_n$

Goal: finding alignments between words and features

---

Aligning words and features

- Use learned conditional probabilities to calculate alignment:
  \[ a(w|f, U, S) = \frac{p(f|w)}{\sum_{w' \in U} p(f|w')} \]

- Normalizing over words: mutual exclusivity effect
Fazly, Alishahi & Stevenson (2010)

- Aligning words and features
  - Use learned conditional probabilities to calculate alignment:
    \[ a(w|f, U, S) = \frac{p(f|w)}{\sum_{w' \in U} p(f|w')} \]
  - Normalizing over words: mutual exclusivity effect

- Updating word-feature associations
  - Word-feature association \( \text{assoc}(w, f) \) can be thought of as alignment-weighted co-occurrence counts
  - \( \text{assoc}(w, f)^t = \text{assoc}(w, f)^{t-1} + a(w|f, U, S) \)
Aligning words and features

- Use learned conditional probabilities to calculate alignment:
  \[ a(w|f, U, S) = \frac{p(f|w)}{\sum_{w' \in U} p(f|w')} \]
- Normalizing over words: mutual exclusivity effect

Updating word-feature associations

- Word-feature association \( \text{assoc}(w, f) \) can be thought of as alignment-weighted co-occurrence counts
  \[ \text{assoc}(w, f)^t = \text{assoc}(w, f)^{t-1} + a(w|f, U, S) \]

Re-calculating \( p(f|w) \) on the basis of \( \text{assoc}(w, f) \)

- Adding smoothing constants for unseen meanings (where \( F \) is the set of all seen features)
  \[ p(f|w) = \frac{\text{assoc}(w,f) + \lambda}{\sum_{f' \in F} \text{assoc}(w,f') + \beta \times \lambda} \]
The data

- What is our input data?
- $U$ is all lemmas in one utterance, e.g. $(doen, daar, maar, in)$
- $S$ is the set of features present in the interval in which $U$ is found.

How to get features from our predicate-argument structures?

- $move(mother,yellow-square-block,in(bucket),on(table))$
- becomes:
  $$\{move,mother,yellow,square,block,bucket,table,in,on\}$$

So an input pair could be:

utterance: doen daar maar in (do there PRT in)
situation: $\{reach,position,floor,on,to,ch,grab,li\}$
Experiments on PAH-game: what do we evaluate

- Words with no clear meaning in our representation (auxiliaries, determiners, many adverbs): don’t evaluate
- 55 lemmas that can be considered meaningful
- Manually annotated for the correct features
- Four types:
  - color and shape terms: rood → red, driehoek → triangular
  - object labels: blok → block
  - spatial terms: op → on, open → \{lid, off, bucket\}
  - verbs halen → \{move, out\}, passen → fit

We evaluate how well the learned $p(f|w)$ at the end fits this ‘golden’ lexicon
Experiments on PAH-game: how do we evaluate

- Two measures, each highlighting a different aspect of the results

**SumProb**  Summed probability: \( \sum_{f \in \text{golden-representation}(w)} p(f|w) \)

**AvePrec**  Average precision: rank the features by \( p(f|w) \), then
\[
\text{AvePrec} = \sum_{k=1}^{n} P(k) \Delta r(k)
\]

where

- \( k \) is the rank
- \( P(k) \) is the number of golden-representation features found up to \( k \), divided by \( k \).
- \( \Delta r(k) \) is the change in recall between \( k - 1 \) and \( k \) (i.e. 1 if a new golden-representation features is found, 0 otherwise).
Experiments on PAH-game: results

- **SP** = Summed Probability
- **AP** = Average Precision

<table>
<thead>
<tr>
<th></th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SP</td>
<td>AP</td>
<td>SP</td>
<td>AP</td>
<td>SP</td>
</tr>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.13</td>
</tr>
</tbody>
</table>
The ‘missing meaning’ problem: windowing

- Suppose the feature is not present at the time of $U$, but some seconds later ... 
- Let the learner pay attention to all intervals between previous and next utterance (flex)
- Or within a fixed window of intervals (e.g. current interval, up to two later; fix)
- So: bigger window of situations covered per utterance
- Sort of socio-pragmatic bootstrapping

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13 0.70</td>
<td>0.05 0.24</td>
<td>0.13 0.25</td>
<td>0.07 0.16</td>
<td>0.09 0.30</td>
</tr>
<tr>
<td>flex</td>
<td>0.10 0.79</td>
<td>0.06 0.23</td>
<td>0.10 0.37</td>
<td>0.08 0.17</td>
<td>0.08 0.34</td>
</tr>
<tr>
<td>fix</td>
<td>0.08 0.73</td>
<td>0.06 0.31</td>
<td>0.10 0.40</td>
<td>0.08 0.21</td>
<td>0.08 0.37</td>
</tr>
</tbody>
</table>
The ‘missing meaning’ problem: adding intentions

- Suppose learners pay attention not only to the current situation, but also what they infer to be the goal of the behavior.
- Goals are game states (in(bucket, block), off(lid, bucket), etc.)
- Goals are inferred using an incrementally trained Naive Bayes Classifier on the basis of the features at $t - 1$.
- Sort of socio-pragmatic bootstrapping

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13 0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.13 0.25</td>
<td>0.07 0.16</td>
</tr>
<tr>
<td>goals</td>
<td>0.11 0.69</td>
<td>0.06</td>
<td>0.29</td>
<td>0.12 0.23</td>
<td>0.07 0.14</td>
</tr>
</tbody>
</table>
The ‘missing words’ problem: adding ghost words

- Suppose some feature is already strongly aligned with a word in the lexicon but not in the utterance
- We can use the strong alignment with that ghost word to make the alignments with the words in the utterance smaller
- Adds a global, probabilistic mutual exclusivity effect
- Let $GW$ be the set of all words seen
- Let $a(w | f, U, S) = \frac{p(f | w)}{\sum_{w' \in U} p(f | w') + \sum_{gw \in GW \land gw \notin U} p(f | gw)}$

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>GW</td>
<td>0.20</td>
<td>0.73</td>
<td>0.07</td>
<td>0.26</td>
<td>0.13</td>
</tr>
</tbody>
</table>
The ‘too much meaning’ problem: weighting by novelty

- Suppose out of all meanings, only some are salient because they’re new.
- Let’s give the new features a high weight and the old ones a low weight.
- $\text{assoc}(w, f)^t = \text{assoc}(w, f)^{t-1} + a(f|w, U, S) \times \text{novelty}(f)$
- Let the novel features be a factor $N$ as likely as old ones,
- $\text{novelty}(f) = \begin{cases} \frac{1}{N \times |\text{novel_features}| + |\text{old_features}|} & \text{if } f \in \text{old_features} \\ \frac{N}{N \times |\text{novel_features}| + |\text{old_features}|} & \text{if } f \in \text{novel_features} \end{cases}$

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13 0.70</td>
<td>0.05 0.24</td>
<td>0.13 0.25</td>
<td>0.07 0.16</td>
<td>0.09 0.30</td>
</tr>
<tr>
<td>$N = 5$</td>
<td>0.08 0.30</td>
<td>0.06 0.22</td>
<td>0.15 0.27</td>
<td>0.09 0.21</td>
<td>0.09 0.24</td>
</tr>
<tr>
<td>$N = 20$</td>
<td>0.05 0.19</td>
<td>0.07 0.20</td>
<td>0.15 0.28</td>
<td>0.10 0.25</td>
<td>0.09 0.22</td>
</tr>
</tbody>
</table>
The ‘too much meaning’ problem: weighting by frequency

- Suppose attention is distributed over all features on the basis of how frequent they are.
- The less frequent, the more salient and vice versa.
- \( \text{assoc}(w, f)^t = \text{assoc}(w, f)^{t-1} + a(w|f, U, S) \times \text{unexpectedness}(f) \)
- \( \text{unexpectedness}(f) = \frac{1}{n(f)} \sum_{f' \in S} \frac{1}{n(f')} \)
- where \( n(f) \) is the frequency of \( f \) in all \( S \) up to time \( t \).

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.09</td>
</tr>
<tr>
<td>freq</td>
<td>0.32</td>
<td>0.68</td>
<td>0.08</td>
<td>0.31</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.33</td>
</tr>
</tbody>
</table>
The ‘too much meaning’ problem: leaving agents out

- The pragmatic situation is very limited
- Therefore the agents child and mother are not salient as they are always present and coincide with the speaker and hearer
- And hence become associated with a lot of words
- Leave them out
- Sort of socio-pragmatic bootstrapping

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>no agt</td>
<td>0.16</td>
<td>0.73</td>
<td>0.06</td>
<td>0.16</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Using distributional information

- Suppose the learner uses the emergent **distributional information** of words
- Frames: word to the left and to the right of $w^8$
- Keep track of an alternative ‘lexicon’ of frames and use that in alignment
  - “go ___ it” will hopefully be associated with verb-like meanings
- Sort of syntactic bootstrapping

\[
a(w|f, fr, U, S) = \frac{p(f|w) + p(f|fr)}{\sum_{w' \in U, fr' = fr(w')} p(f|w') + p(f|fr')}
\]

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>frames</td>
<td>0.10</td>
<td>0.70</td>
<td>0.05</td>
<td>0.23</td>
<td>0.12</td>
</tr>
</tbody>
</table>

An overview: what helps, what doesn’t

- Model does not learn that well from the data
- It is to be seen if other models do: problem seems inherent in the data
- But also tells us something about the task the learner faces
- Main (global) positive effects:
  - a wider window into the future (0 : 2 seems to work best)
  - weighting by inverse frequency
  - adding ghost words
  - leaving agents out
An overview: what helps, what doesn’t

<table>
<thead>
<tr>
<th>model</th>
<th>color/shape</th>
<th>object</th>
<th>spatial</th>
<th>verbs</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>0.13</td>
<td>0.70</td>
<td>0.05</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>4-best</td>
<td>0.22</td>
<td>0.77</td>
<td>0.09</td>
<td>0.28</td>
<td>0.23</td>
</tr>
</tbody>
</table>

- Only slightly better in Average Precisions than the windowing approach (0.37 vs. 0.38)
- But much better in Summed Probability (0.08 vs. 0.16)
We can get some improvement using low-level cues:
- SumProb from 0.09 to 0.16
- AvePrec from 0.30 to 0.38

Continuing search for other cues (prosody?)
Also general conclusion: the ‘cross-situationality’ of this data is limited
But perhaps also: aligning single words with features might not be realistic
- `ga @m d@r m@ in doen`
  - `go it there PRT in do.INF`
  - ‘go put it in there’
- has a fixed part, recurring over tens of utterances
- Variable are: `in doen` (put in), `in stoppen` (put in) `uit halen` (take out), `op zetten` (put on).
- Can this information somehow be exploited?
Thank you!