Learning relational meanings from situated caregiver-child interaction

A computational approach

Barend Beekhuizen[1], Afsaneh Fazly[2], Aida Nematzadeh[2] & Suzanne Stevenston[2]

[1]Leiden University [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?

Learning word meaning

- 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?

Relational meaning

- 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)

[1] Gentner (1978) 'On relational meaning: The acquisition of verb meaning'. *Child Development* **49**:988–998
Gleitman (1990): 'Sources of verb meanings'. *Language Acquisition* **1**: 3–55

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.

[4] Behrend, D. A., & J. Scofeld (2006). 'Verbs, Actions, and Intentions'. In: K. Hirsh-Pasek & R. M. Golinkoff (eds.). *Action Meets Word. How Children Learn Verbs*, p. 286–307

[5] Poulin-Dubois, D., & J. N. Forbes (2006). 'Word, Intention, and Action: A Two-Tiered Model of Action Word Learning'. In K. Hirsh-Pasek & R. M. Golinkoff (eds.), *Action Meets Word. How Children Learn Verbs*, p. 262–285

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).[6]
- Get a more detailed picture of the problems.
- Then: using computational modeling techniques to test some of the hypothesized solutions

[6] Frank, M.C., N.D. Goodman & J.B. Tenenbaum (2008). 'A Bayesian Framework for Cross-Situational Word-Learning'. *Advances in Neural Information Processing Systems*, **20**, 18
Medina, T. N., J. Snedeker, J.C. Trueswell & L.R. Gleitman (2011). 'How words can and cannot be learned by observation'. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 9014–9

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

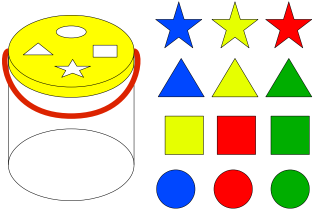

Figure: The toy and the twelve blocks

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

Fazly, Alishahi & Stevenson (2010)

- Starting point: model of Fazly, Alishahi & Stevenson (2010); FAS[7]
- 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

[7] Fazly, A., A. Alishahi & S.Stevenson, (2010). 'A probabilistic computational model of cross-situational word learning'. *Cognitive science*, **34**, 1017–1063.

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

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 $assoc(w, f)$ can be thought of as alignment-weighted co-occurrence counts
  - $assoc(w, f)^t = assoc(w, f)^{t-1} + a(w|f, U, S)$

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 $assoc(w, f)$ can be thought of as alignment-weighted co-occurrence counts
  - $assoc(w, f)^t = assoc(w, f)^{t-1} + a(w|f, U, S)$
- Re-calculating $p(f|w)$ on the basis of $assoc(w, f)$
  - Adding smoothing constants for unseen meanings (where F is the set of all seen features)
  - $p(f|w) = \frac{assoc(w,f)+\lambda}{\sum_{f' \in F} assoc(w,f')+\beta\times\lambda}$

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* $\rightarrow$ `red`, *driehoek* $\rightarrow$ `triangular`
  - object labels: *blok* $\rightarrow$ `block`
  - spatial terms: *op* $\rightarrow$ `on`, *open* $\rightarrow$ {`lid`,`off`,`bucket`}
  - verbs *halen* $\rightarrow$ {`move`,`out`}, *passen* $\rightarrow$ `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 $AvePrec = \sum_{k=1}^{n} P(k)\Delta r(k)$

- 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

	color/shape		object		spatial		verbs		**total**	
	SP	AP	SP	AP	SP	AP	SP	AP	SP	AP
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**

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

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
flex	0.10	0.79	0.06	0.23	0.10	0.37	0.08	0.17	**0.08**	**0.34**
fix	0.08	0.73	0.06	0.31	0.10	0.40	0.08	0.21	**0.08**	**0.37**

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

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
goals	0.11	0.69	0.06	0.29	0.12	0.23	0.07	0.14	**0.08**	**0.31**

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
- $a(w|f,U,S) = \frac{p(f|w)}{\sum_{w' \in U} p(f|w') + \sum_{gw \in GW \wedge gw \notin U} p(f|gw)}$

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
GW	0.20	0.73	0.07	0.26	0.13	0.26	0.07	0.17	**0.10**	**0.32**

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
- $assoc(w, f)^t = assoc(w, f)^{t-1} + a(f|w, U, S) \times novelty(f)$
- Let the novel features be a factor N as likely as old ones,
- $novelty(f) = \frac{1}{N \times |novel\_features| + |old\_features|}$ if $f \in old\_features$

  $novelty(f) = \frac{N}{N \times |novel\_features| + |old\_features|}$ if $f \in novel\_features$

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
$N = 5$	0.08	0.30	0.06	0.22	0.15	0.27	0.09	0.21	**0.09**	**0.24**
$N = 20$	0.05	0.19	0.07	0.20	0.15	0.28	0.10	0.25	**0.09**	**0.22**

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
- $assoc(w, f)^t =$
  $assoc(w, f)^{t-1} + a(w|f, U, S) \times unexpectedness(f)$
- $unexpectedness(f) = \frac{\frac{1}{n(f)}}{\sum_{f' \in S} \frac{1}{n(f')}}$
- where $n(f)$ is the frequency of f in all S up to time t.

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
freq	0.32	0.68	0.08	0.25	0.14	0.31	0.14	0.23	**0.15**	**0.33**

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

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
no agt	0.16	0.73	0.06	0.26	0.16	0.26	0.08	0.18	**0.10**	**0.32**

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

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
frames	0.10	0.70	0.05	0.23	0.12	0.26	0.07	0.17	**0.08**	**0.30**

[8] Mintz, T. H., E.L Newport & T.G. Bever (2002). 'The distributional structure of grammatical categories in speech to young children'. *Cognitive Science* **26**, 393–424

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

model	color/shape		object		spatial		verbs		**total**	
basic	0.13	0.70	0.05	0.24	0.13	0.25	0.07	0.16	**0.09**	**0.30**
4-best	0.22	0.77	0.09	0.28	0.23	0.41	0.17	0.24	**0.16**	**0.38**

- 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)

Whither?

- 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!