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

ICLC, 25 June 2013

Introduction

Topic

Cognitive models of acquiring word-meaning mappings

Goals

1. methodological issues: Discuss sources of semantic data for models and present a new one
2. providing a baseline: Explore the behavior of a basic word-learning model on this data
3. extending the model: Show how we can add 'modules' to the model

- Cross-situational learning → computational models
- Input: utterances and situations (source: synthetic or video)

- Cross-situational learning $\rightarrow$ computational models
- Input: utterances and situations (source: synthetic or video)

Goal #1

Provide situational descriptions (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.

- Cross-situational learning → computational models
- Input: utterances and situations (source: synthetic or video)

Goal #1

Provide situational descriptions (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.

- 32 dyads (child 16mo, ± 5 min. each) playing game.
- 175 minutes of material, 7842 word tokens, 2492 utterances.
- Situational coding. For every interval of 3 seconds, code:
  - simple behavior (`grab`,`move`,`position`,`letgo`),
  - changes in spatial relations (`in`,`on`,`out`,`off`,`match`),
  - objects (`block`,`bucket`,`mother`,`table`)
  - properties (`triangular`,`square`,`red`,`blue`)

- Cross-situational learning → computational models
- Input: utterances and situations (source: synthetic or video)

Goal #1

Provide situational descriptions (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.

- 32 dyads (child 16mo, $\pm$ 5 min. each) playing game.
- 175 minutes of material, 7842 word tokens, 2492 utterances.
- Situational coding. For every interval of 3 seconds, code:
  - simple behavior (`grab`,`move`,`position`,`letgo`),
  - changes in spatial relations (`in`,`on`,`out`,`off`,`match`),
  - objects (`block`,`bucket`,`mother`,`table`)
  - properties (`triangular`,`square`,`red`,`blue`)
- Structured: `grab(mother,(red,square,block))`
- High intra- & interannotator agreement (almost all $\kappa > 0.8$)

Example

time	type	coding/transcription
0m0s	**situation**	
	language	een. nou jij een.
	translation	"One. Now you try one."
0m3s	**situation**	position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro)
	language	nee daar.
	translation	"No, there."
0m6s	**situation**	point(mother, ho-tr, child) position(child, b-ye-tr, near(b-ye-tr, ho-ro)) mismatch(b-ye-tr, ho-ro)
	language	nee lieverd hier past ie niet.
	translation	"No sweetie, it won't fit in here."

Acquiring lexical meaning

Goal #2

Setting a baseline: how well does a word-learning model like Fazly et al. 2010 (FAS10) perform on this data?

Acquiring lexical meaning

Goal #2

Setting a baseline: how well does a word-learning model like Fazly et al. 2010 (FAS10) perform on this data?

- FAS10: incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$

Acquiring lexical meaning

Goal #2

Setting a baseline: how well does a word-learning model like Fazly et al. 2010 (FAS10) perform on this data?

- FAS10: incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$
- Data preparation
  - Representations are structured, so flatten them: $\texttt{grab(mother,(red,square,block))} \rightarrow \{\texttt{grab,mother,red,square,block}\}$
  - Take the set of all flattened representations of the situations occurring in the interval in which the utterance was produced.
  - We used lemma representations for the words

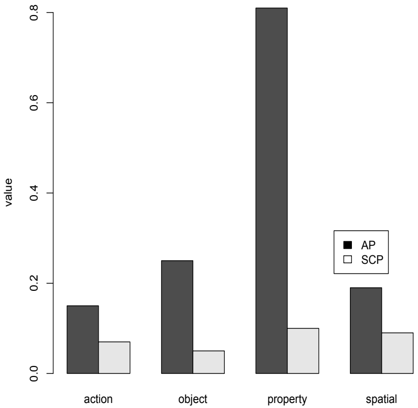
Baseline experiment: evaluation

- No golden lexicon, so hand-built one for 'meaningful' words ($n = 41$):
    - Object labels: *blok* meaning `block`
    - Properties: *rood* meaning `red`
    - Spatial relations: *op* meaning `on`
    - Actions: *passen* meaning `match`, *stoppen* meaning {`move`,`in`}

Baseline experiment: evaluation

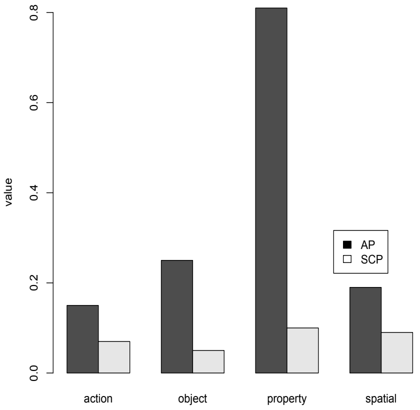
- No golden lexicon, so hand-built one for ‘meaningful’ words ($n = 41$):
  - Object labels: *blok* meaning `block`
  - Properties: *rood* meaning `red`
  - Spatial relations: *op* meaning `on`
  - Actions: *passen* meaning `match`, *stoppen* meaning {`move`,`in`}
- Two (partially complementary) measures:
  - Summed Conditional Probability (*SCP*): how much probability mass is assigned to the true meanings given a word?
  - Average Precision (*AP*): how are the true meanings ranked (on conditional probability) w.r.t. the other meanings.

Results



- *SCP* not very peaky
- *AP* (ranking): good for properties, rather bad for other classes.
- No model dependence.

Results



- *SCP* not very peaky
- *AP* (ranking): good for properties, rather bad for other classes.
- No model dependence.
- Relational meanings hard to glean from situation alone. Why?
  1. True meaning absent from *S*
  2. Foil features structurally present in
  3. True meaning also present in many other *S*s
- In general: situations look a lot like each other, unlike 'synthesized' semantics (cf. Matusevych et al. 2013)

Goal #3

Exploring known biases/mechanisms

added bias/mechanism	prop.	object	spatial	actions
INTENTION				
increasing temporal scope	$=$	$\uparrow$	$\uparrow$	$\uparrow$
attention to own behavior	$=$	$\downarrow$	$\downarrow$	$\uparrow$
attention to mother's behavior	$\uparrow$	$\downarrow$	$\downarrow$	$=$
ATTENTION				
only take novel features	$\downarrow$	$\downarrow$	$\downarrow$	$\uparrow$
more weight to novel features	$\downarrow$	$\downarrow$	$\uparrow$	$\uparrow$
more weight to rarer features	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
more weight to expected features	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$=$
LINGUISTIC STRUCTURE				
using parts of speech	$=$	$\downarrow$	$=$	$=$
Mintz' frequent frames	$\downarrow$	$\downarrow$	$=$	$\uparrow$

1. Data issues for word learning models
   - problems with synthesizing methods and typical video-based approaches
   - creation of a situational corpus

1. Data issues for word learning models
    - problems with synthesizing methods and typical video-based approaches
    - creation of a situational corpus
2. Setting a baseline using FAS10
    - properties $>$ object labels $>$ spatial & behavioral meaning
    - other methods underestimate noise & uncertainty in actual data

1. Data issues for word learning models
    - problems with synthesizing methods and typical video-based approaches
    - creation of a situational corpus
2. Setting a baseline using FAS10
    - properties > object labels > spatial & behavioral meaning
    - other methods underestimate noise & uncertainty in actual data
3. Exploring other mechanisms
    - method to evaluate their contribution
    - what works:
        - attention to rare events,
        - increasing temporal scope,
        - adding words from previous utterances
    - other mechanisms are mixed: e.g. good for verbs, bad for rest

FAS10

- Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w' \in U^{(t)}} p^{(t-1)}(f|w')} \tag{1}$$

FAS10

- Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w' \in U^{(t)}} p^{(t-1)}(f|w')} \quad (1)$$

- Updating the association score (initialized at 0):

$$\mathrm{assoc}^{(t)}(w, f) = \mathrm{assoc}^{(t-1)}(w, f) + a(w|f, U^{(t)}, S^{(t)}) \quad (2)$$

FAS10

- Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum\limits_{w' \in U^{(t)}} p^{(t-1)}(f|w')} \quad (1)$$

- Updating the association score (initialized at 0):

$$\text{assoc}^{(t)}(w, f) = \text{assoc}^{(t-1)}(w, f) + a(w|f, U^{(t)}, S^{(t)}) \quad (2)$$

- Recalculating the conditional probabilities:

$$p^{(t)}(f|w) = \frac{\text{assoc}^{(t)}(w, f) + \lambda}{\sum\limits_{f' \in F} \text{assoc}^{(t)}(w, f') + \beta \times \lambda} \quad (3)$$