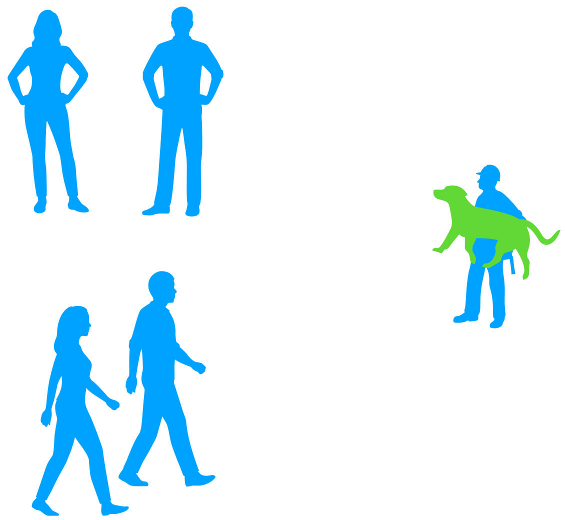
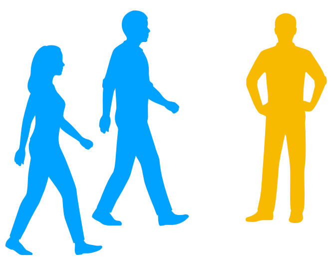
A practical compositional semantics for situated interaction

David Schlangen
Universität Bielefeld
CARLA Workshop, August 2018

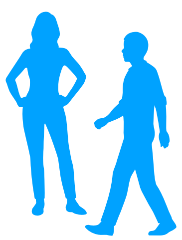
http://dsg-bielefeld.de/talks/carla-2018.html



A: *Look at the dog!*
B: *I know. Isn't it cute?*

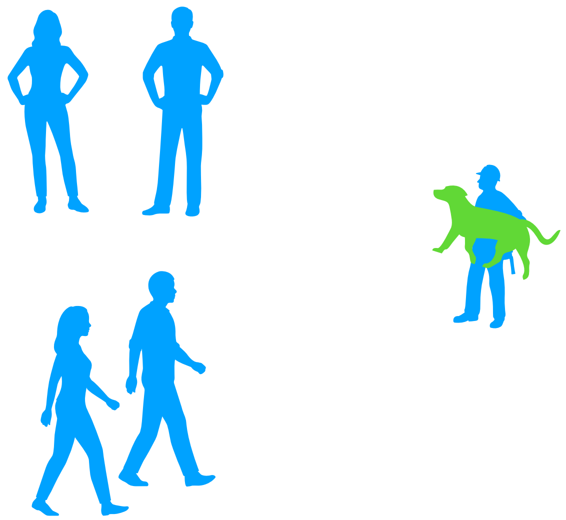


A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*

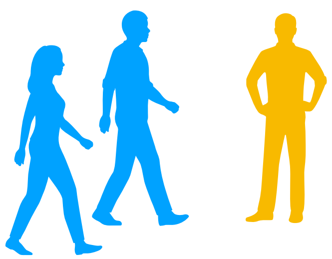


A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*

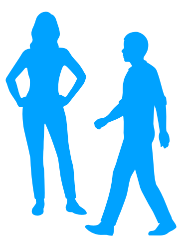
http://dsg-bielefeld.de/talks/carla-2018.html



A: *Look at the dog!*
B: *I know. Isn't it cute?*

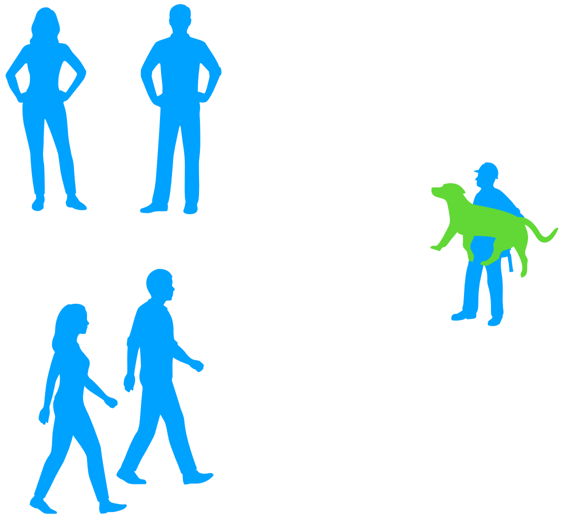


A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*

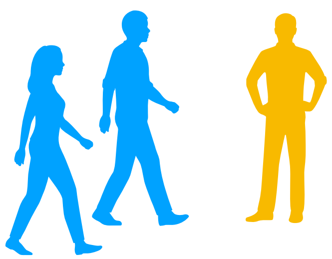


A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*

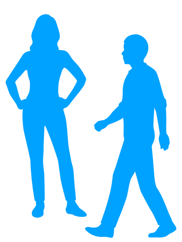
http://dsg-bielefeld.de/talks/carla-2018.html



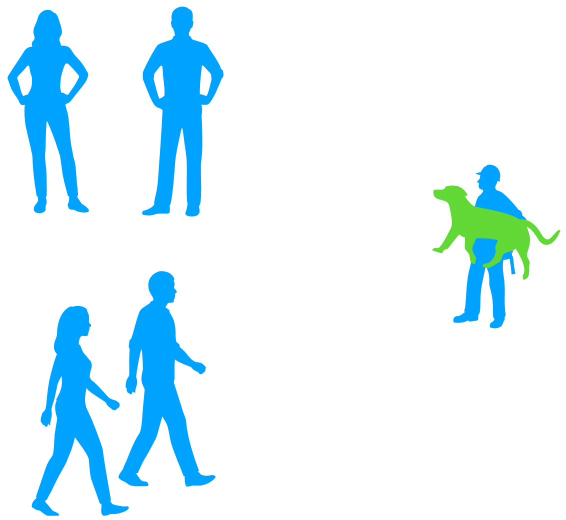
A: *Look at the dog!*
B: *I know. Isn't it cute?*



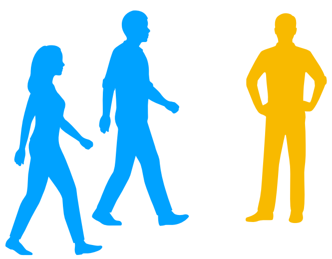
A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*



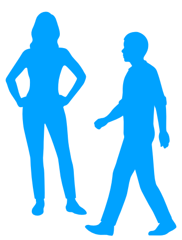
A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*



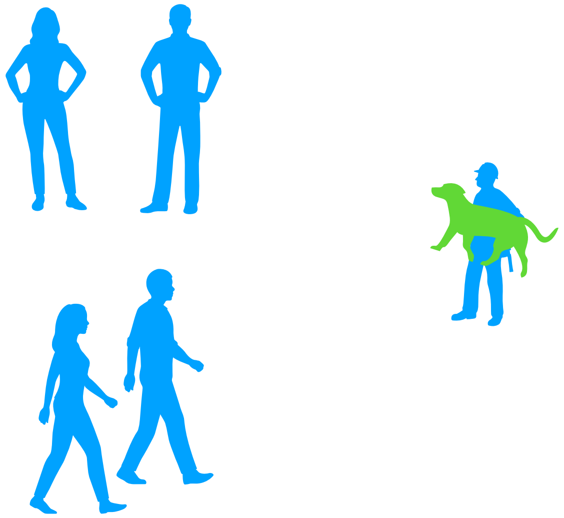
A: *Look at the dog!*
B: *I know. Isn't it cute?*



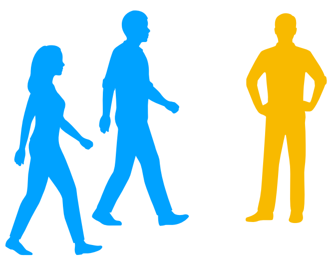
A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*



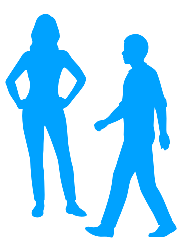
A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*



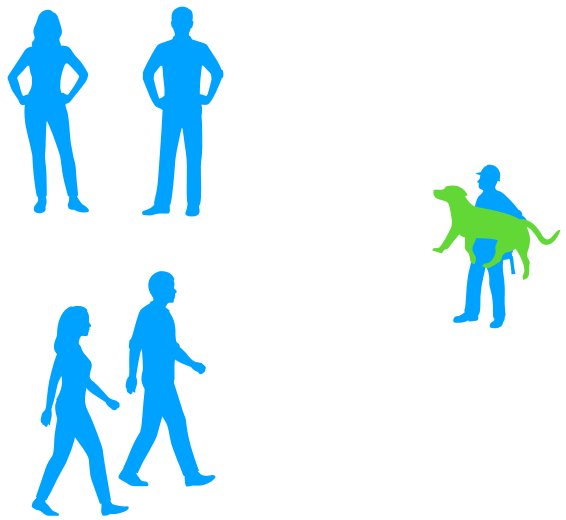
A: *Look at the dog!*
B: *I know. Isn't it cute?*



A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*

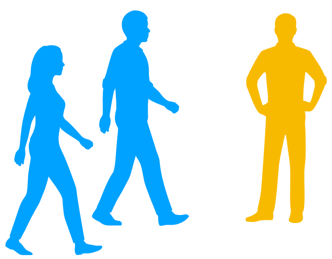


A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*



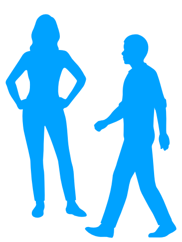
A: *Look at the dog!*
B: *I know. Isn't it cute?*

exophoric reference



A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*

co-reference

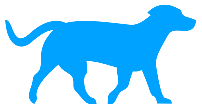


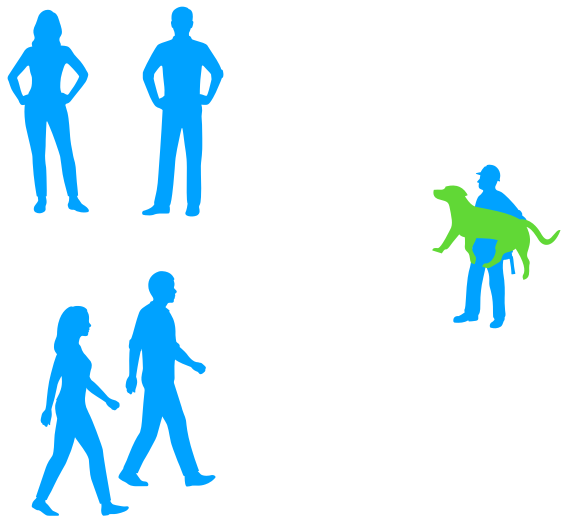
A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*

meta-semantic interaction

trust

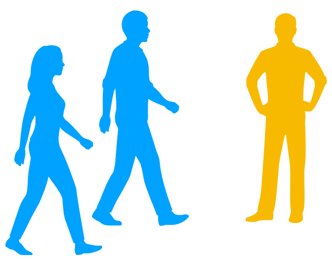
Desiderata for use of concepts in situated interaction

- language-to-world [“dog” to ]
- language-to-language [“poodle” to “dog”]
- negotiable, update-able
- language-to-expert



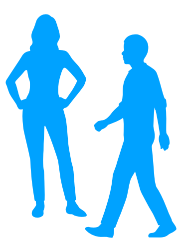
A: *Look at the dog!*
B: *I know. Isn't it cute?*

learning from demonstration



A: *We just saw a man carrying a dog.*
B: *The cutest poodle ever!*

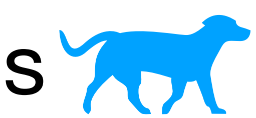
learning from definition



A: *I don't think that was a poodle. It was too tall. I think it was a labradoodle.*
B: *Oh. I guess you're right. You're the expert here.*

learning from syntactic contexts

Desiderata for learning of concepts from situated interaction

- from instance demonstration [“dog” referring to this ]
- from being given facts [“a labradoodle is a cross between labrador retriever and poodle”]
- from overhearing contexts

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
  - Learning to resolve reference, based on definitions
  - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

http://dsg-bielefeld.de/talks/carla-2018.html

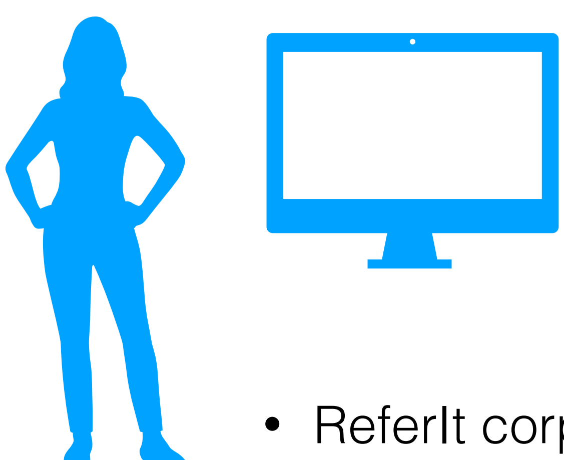
word-to-world

- Very straightforward:
  - Given corpus of referring expression + referred object in image, …
  - …train classifier for all words, predicting how well word fits object.
- “Words as classifiers” approach; IWCS 2015, ACL 2016, ACL 2017.
- (See also [Harnad 1990, Roy *et al.* 2002, Siebert & Schlangen 2008, Larsson 2013].)

word-to-world

- Very straightforward:
  - Given corpus of referring expression + referred object in image, …
  - …train classifier for all words, predicting how well word fits object.
- “Words as classifiers” approach; IWCS 2015, ACL 2016, ACL 2017
- (See also [Harnad 1990, Roy *et al.* 2002, Siebert & Schlangen 2008, Larsson 2013].)

A Corpus of Referential Interactions

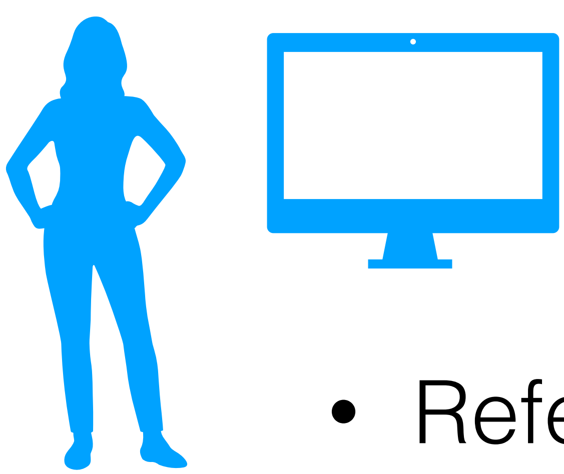


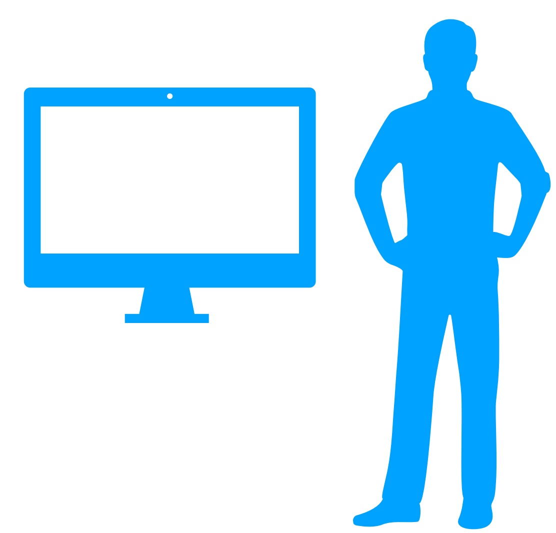
"Person left" √

- ReferIt corpus (Kazemzadeh *et al.* 2014): 20k images (SAIAPR, [Escalante *et al.* 2010]), 120k referring expressions
- MSCOCO (Lin *et al.* 2014): 27k images, 100k region descriptions (Mao *et al.* 2015) + 140k referring expressions (Berg *et al.* 2015) + 140k (non-positional) ref exp (Yu *et al.* 2016)

A Corpus of Referential Interactions

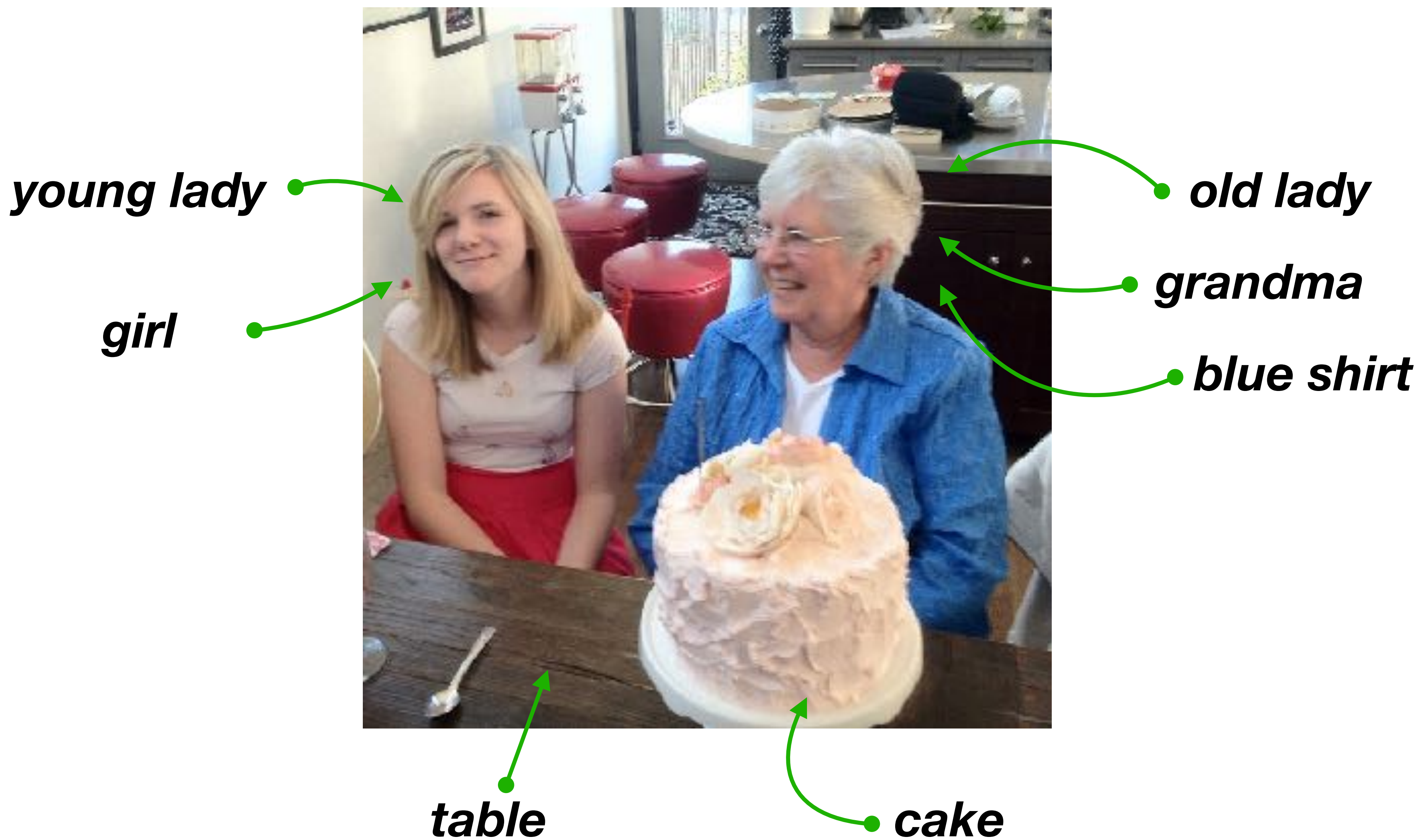
“Person left” √





- Referring expressions, not labels!
  - No closed-world assumption.
  - No pre-conceived tagset.

A Corpus of Referential Interactions



word-to-world

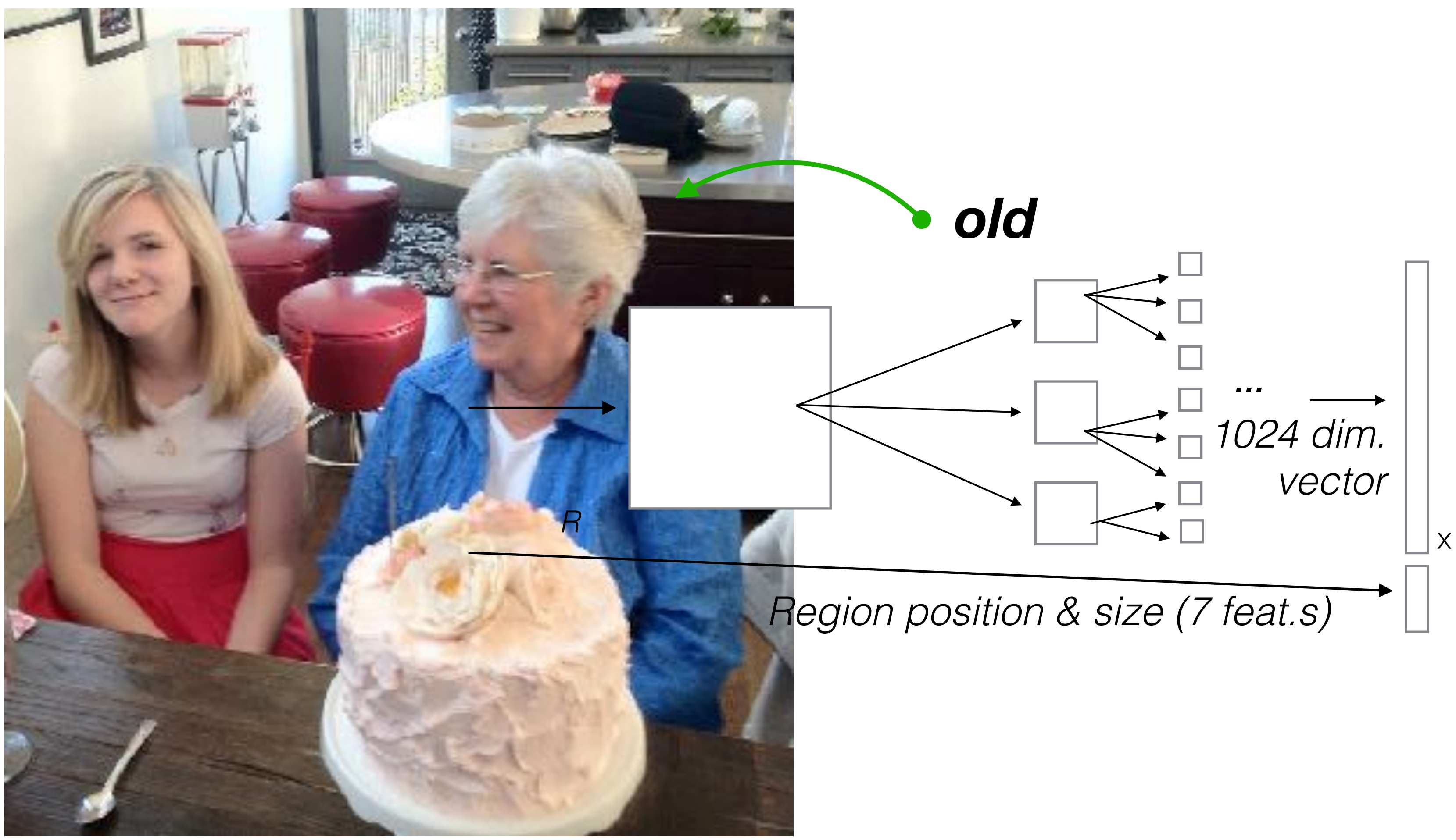
- Very straightforward:
  - Given corpus of referring expression + referred object in image, …
  - …train classifier for all words, predicting how well word fits object.
- “Words as classifiers” approach; IWCS 2015, ACL 2016

Acquiring Referential Competence

old lady

First assumption:
We can learn words independently.

Acquiring Referential Competence



$$\sigma(\ \cdot\ ;\ \Theta_{old})$$

First assumption:
We can learn words independently.

Extract visual features.
Pre-trained CNN (GoogLeNet; Szegedy et al., 2015) + positional features.

Randomly sample other regions as negative instances.

Train logistic regression classifier for current word.
(L1-regulated, cross-entropy, SGD.)

Training

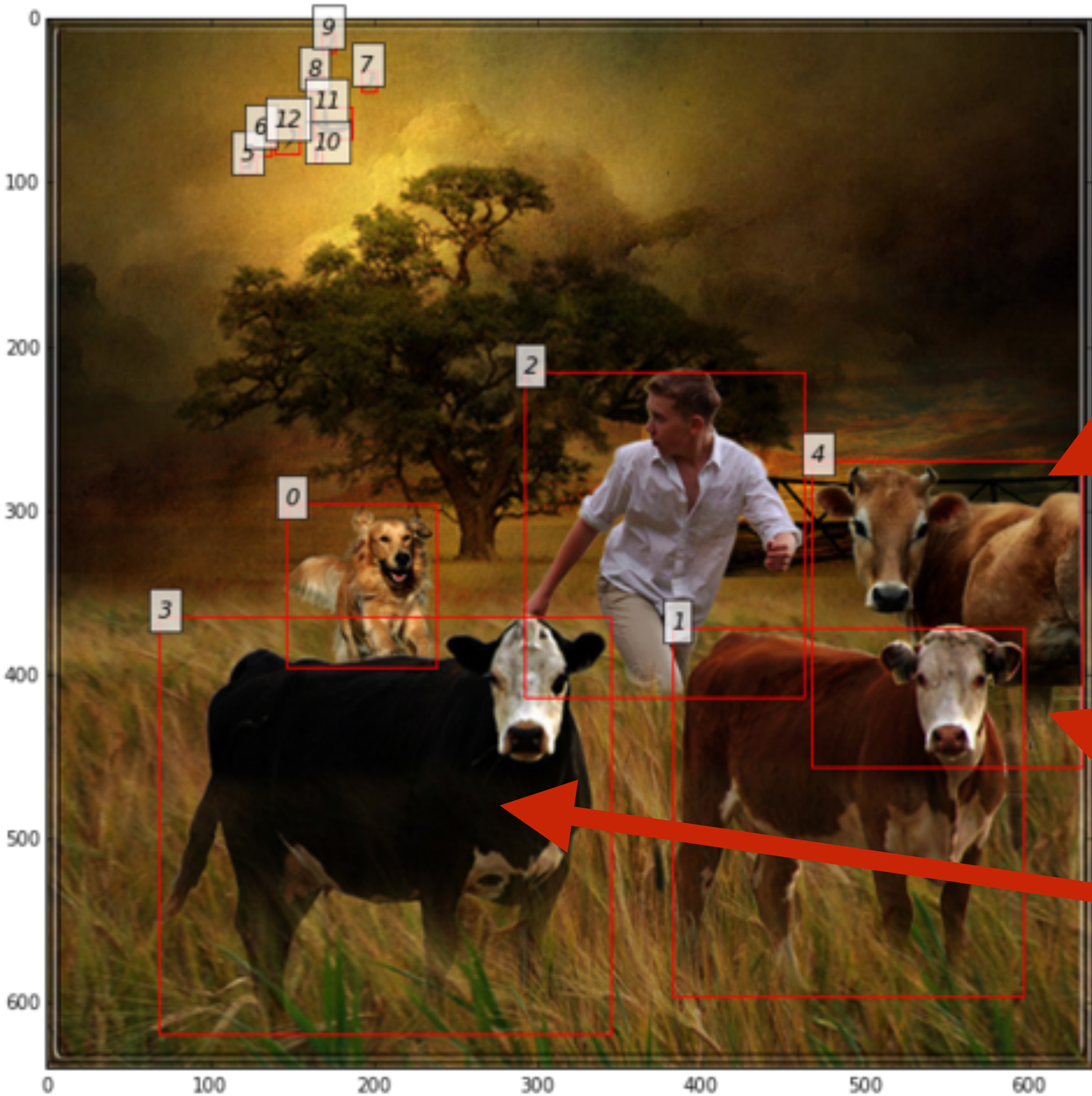
Guy with white shirt

Training

Guy with white shirt

¬Guy ¬with ¬white ¬shirt

Training



Cow
right

¬Cow
¬right

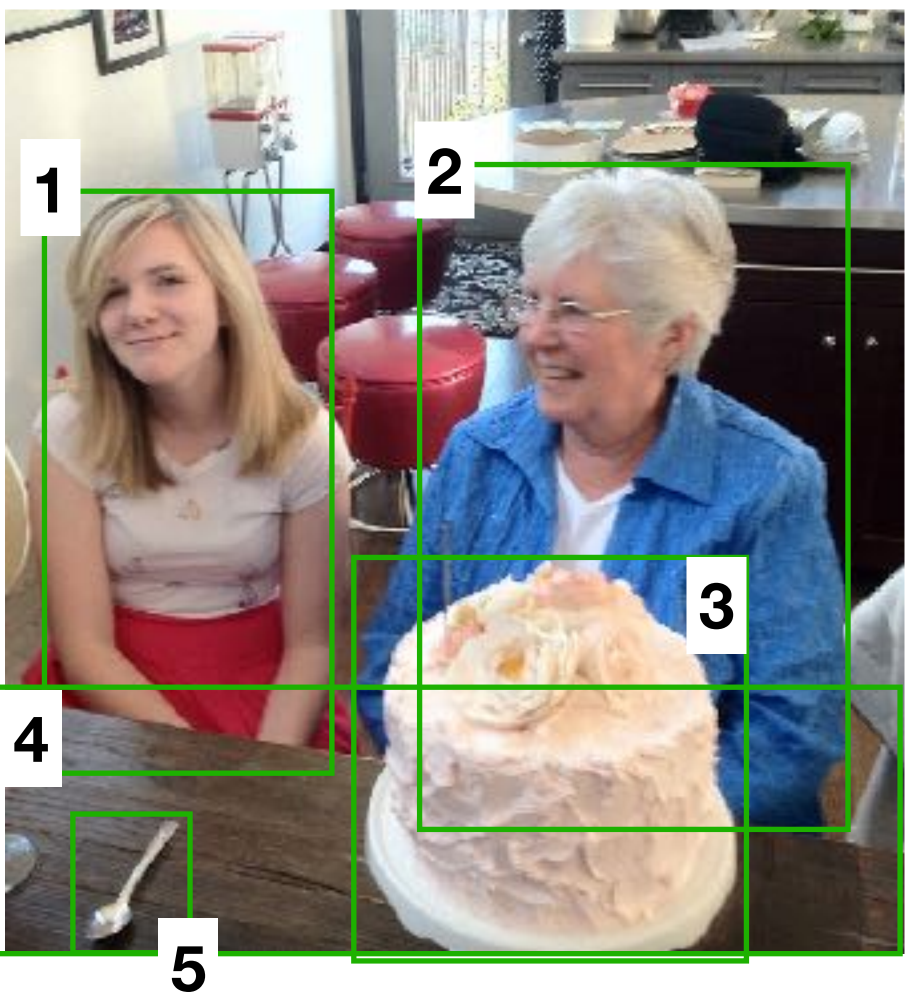
Second assumption:
If a property has not been mentioned when referring to an object, it doesn't have it.

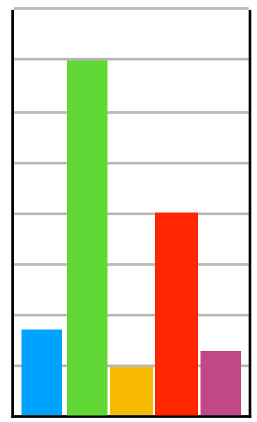
Acquiring Referential Competence

old

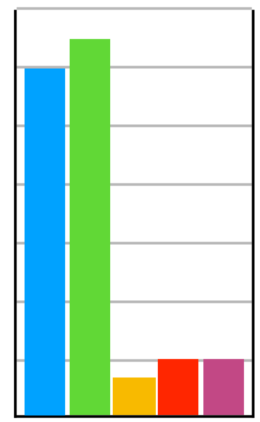
Resolving References

the old lady

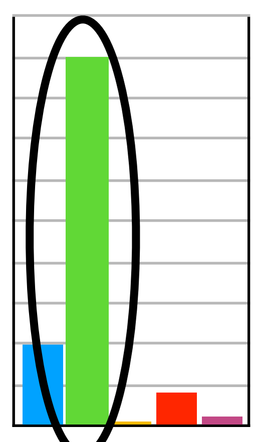




old



lady



old⊙lady

the old lady

$$the_x\,[\,old(x) \land lady(x)\,]$$

$$D = \{\,o_1, o_2, o_3, o_4, o_5\,\}$$

$$\cancel{F(old) = \{\,o_2, o_4\,\}}$$

$$\cancel{F(old) = [\,0, 1, 0, 1, 0\,]}$$

~~**∧ is intersection**~~

$$F(old) = \lambda x.\sigma(x\,;\theta_{old})$$

$$F(old) = [\,0.1, 0.7, 0.2, 0.4, 0.1\,]$$

∧ is multiplication

***the* is argmax**

Results

	%tst	acc	mrr	arc	>0	acc
REFERIT	1.00	0.65	0.79	0.89	0.97	0.67
REFERIT; NR	0.86	0.68	0.82	0.91	0.9	**0.71**
(Hu et al., 2015)	–	0.73	–	–	–	–
REFCOCO	1.00	0.61	0.77	0.91	0.98	0.62
REFCOCO; NR	0.94	0.63	0.78	0.92	0.9	**0.64**
(Mao et al., 2015)	–	0.70	–	–	–	–
GREXP	1.00	0.43	0.65	0.86	1.00	0.43
GREXP; NR	0.82	0.45	0.67	0.88	1.0	**0.45**
(Mao et al., 2015)	–	0.61	–	–	–	–

Results, full model

	RP@1	RP@10	rnd
REFERIT	0.09	0.24	0.03
REFERIT; NR	**0.10**	0.26	0.03
(Hu et al., 2015)	0.18	0.45	
REFCOCO	0.52	–	0.17
REFCOCO; NR	**0.54**	–	0.17
(Mao et al., 2015)	0.52		
GREXP	0.36	–	0.16
GREXP; NR	**0.37**	–	0.17
(Mao et al., 2015)	0.45		

Region Proposals

	nopos	pos	full	top20
RI	0.53	0.60	0.65	0.46
RI; NR	0.56	0.62	0.68	0.48
RC	0.44	0.55	0.61	0.52
RC; NR	0.45	0.57	0.63	0.53

Feature Ablation

(Schlangen, Zarrieß, Kennington; ACL 2016)

not state of the art, why bother?

- **not end-to-end, is inspectable, so allows us to play around with word models**
- **“dialogue ready”, as it is triply incremental:**
    - **open vocab set, can always learn new words**
    - **can always continue to learn model of a word**
    - **application is incremental**

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
  - Learning to resolve reference, based on definitions
  - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
    - Learning to resolve reference, based on definitions
    - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

Zero-shot learning from definitions

- Zero-shot learning in CV: take information from a different source and use it to make visual categorisation decisions. (E.g., Lampert *et al.* 2009)

- Here, again very straightforward. Replace term with its definition and resolve in the normal way (applying word classifiers in definiens).

    - E.g., replace “SUV” (for which no visual classifier exists) with “large car”

Induce structure in lexicon

- Turn classifiers (trained from pairings of word and image of referent) into vectors in metric space.
- Use distance as indicator of semantic similarity.
- Use usual tricks to infer relations. (E.g., hypernym should have higher entropy. [Kiela *et al.* 2015].)
- Results: Reproducing similarity judgements kind of works. Interesting errors. (E.g., predicts that “scarf” is a type of “woman”.)

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
  - Learning to resolve reference, based on definitions
  - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
  - Learning to resolve reference, based on definitions
  - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

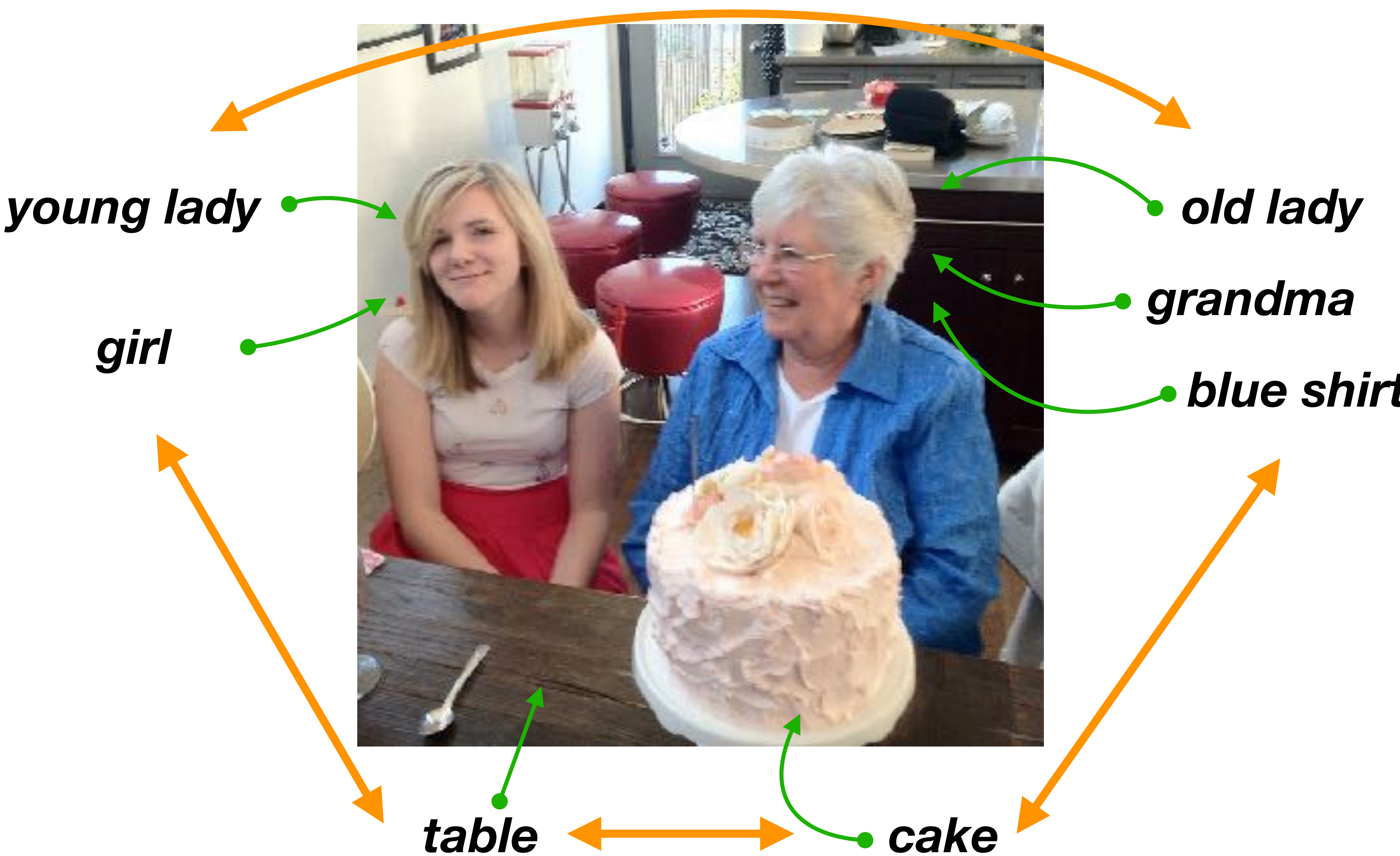
Word representations from situated contexts

Problem: word2vec etc. predict as very similar e.g.

- *left* and *right*
- *red* and *green*

Which is correct in some ways, and unhelpful in others.

Word representations from situated contexts



embeddings, from different kinds of context:

- **ref.exp. as sentence, whole corpus**
- **co-referential exp. as context**
- **situation as context**

(Zarrieß & Schlangen, EMNLP 2017)

Evaluating Derived Concept Relations

Similarity / Relatedness / Compatibility

Model	MEN	SemSim	VisSim	Compatibility
w2v_ref	0.669	**0.687**	**0.580**	**0.251**
w2v_den	**0.765**	0.651	0.570	0.164
w2v_sit	0.586	0.515	0.409	0.166
baronimod	0.785	0.704	0.594	0.241
vis_av	0.523	0.526	0.486	0.287
wac_int	-0.373	-0.339	-0.294	-0.076
wac_den	-0.593	-0.615	-0.536	**-0.288**
wac_resp	**0.634**	**0.656**	**0.574**	0.276

(Baroni *et al.* 2014) CBOW, 400dim

(Bruni *et al.* 2012)
372 out of 3,000

(Silberer & Lapata 2014)
721 out of 7,577

(Kruszewski & Baroni 2015)
1,859 out of 17,973

(Zarrieß & Schlangen, EMNLP 2017)

Predicting Incompatible Modifiers

Similar according to linguistic context (whole corpus), dissimilar according to referential context (refer to same entity).

man

left $\longleftrightarrow$ right

young $\longleftrightarrow$ old

old $\longleftrightarrow$ shirtless

shirt

plaid $\longleftrightarrow$ green

red $\longleftrightarrow$ gray

blue $\longleftrightarrow$ yellow

elephant

closest $\longleftrightarrow$ back

big $\longleftrightarrow$ baby

adult $\longleftrightarrow$ smaller

Overview

- Motivation, Goals, Desiderata
- Learning to connect words and images, and referring expressions and images
- Crossover:
  - Learning to resolve reference, based on definitions
  - Learning to connect concepts, based on exposure to instances
- Learning connections between concepts from situated contexts
- Outlook

Desiderata

- Use:
  - to pick out objects in the world / language-to-world
  - to pick out objects in the discourse / language-to-language
  - to be object of discussion
  - to reside in others
- Learning:
  - from demonstration

words-as-classifiers (WAC)

semantic similarity from WACs & from structured contexts

What notion of “concept” is this? … Conceptual pluralism…
Follows Lewis’s advice [Partee 1995 / Lewis 1970], “meaning is what meaning does”…

Thank you!

http://dsg-bielefeld.de/talks/carla-2018.html

Joint work with Casey Kennington & Sina Zarrieß, with input from the whole Dialogue Systems Group Bielefeld.

Universität Bielefeld

dialogue systems group [unibi]

CITEC
Cognitive Interaction Technology
Cluster of Excellence
Bielefeld University

References

- Zarrieß S, Schlangen D. Deriving continous grounded meaning representations from referentially structured multimodal contexts. In: Proceedings of EMNLP 2017 – Short Papers. 2017. PDF
- Zarrieß S, Schlangen D. Refer-iTTS: A System for Referring in Spoken Installments to Objects in Real-World Images. In: Proceedings of INLG 2017 (demo papers). 2017. PDF
- Zarrieß S, Schlangen D. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In: Proceedings of 55th annual meeting of the Association for Computational Linguistics (ACL). Vancouver. 2017 PDF
- Zarrieß S, Schlangen D. Is this a Child, a Girl, or a Car? Exploring the Contribution of Distributional Similarity to Learning Referential Word Meanings. In: Short Papers – Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL). 2017 PDF
- Schlangen D, Zarrieß S, Kennington C. **Resolving References to Objects in Photographs using the Words-As-Classifiers Model**. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin; 2016. PDF
- Manuvinakurike R, Kennington C, DeVault D, Schlangen D. Real-Time Understanding of Complex Discriminative Scene Descriptions. In: Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue. 2016. PDF
- Zarrieß S, Schlangen D. Towards Generating Colour Terms for Referents in Photographs: Prefer the Expected or the Unexpected? In: Proceedings of the 9th International Natural Language Generation conference. Edinburgh, UK: Association for Computational Linguistics; 2016: 246–255. PDF
- Zarrieß S, Schlangen D. Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). 2016. PDF
- Schlangen D. **Grounding, Justification, Adaptation: Towards Machines That Mean What They Say**. In: Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem). 2016. PDF
- Kennington C, Schlangen D. **Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution**. In: Proceedings of the Conference for the Association for Computational Linguistics (ACL). Association for Computational Linguistics; 2015: 292–301. PDF