Language, Action, and Perception

Simon Dobnik

31 European Summer School in Logic,
Language and Information
12 August 2019, Riga Latvia

CLASP centre for linguistic theory and studies in probability

Welcome to the course

What is this course about?

UNIVERSITY OF GOTHENBURG

- Computational modelling of language, action and perception in relation to image classification and situated dialogue agents

What is this course about?

UNIVERSITY OF GOTHENBURG

- Computational modelling of language, action and perception in relation to image classification and situated dialogue agents
- Relates to:
  - linguistics
  - experimental psychology
  - computer science
  - computer vision
  - robotics
  - artificial intelligence

We will discuss three kinds of topics

UNIVERSITY OF GOTHENBURG

- Linguistics and psychology: how humans connect language, spatial perception, action?
- Formal computational systems: what kind of models and algorithms do we employ?
- Applications: what kind of problems do we want to solve?

Relation to robotics

UNIVERSITY OF GOTHENBURG

- Spatial cognition and action represent the core of human cognition and behaviour.
- A robot that can make sense of the world and interact with humans is very useful: navigation systems, assistants to people with disabilities, robots on rescue missions, just for fun, etc.
- Having access to robot' sensors and actuators can give us a theoretical insight into language, spatial perception and action.

Relation to computer vision

UNIVERSITY OF GOTHENBURG

- Social media includes text, images and videos
- Visual information closely linked to textual data, e.g. a newspaper article or a Facebook post
- Can we make sense of it?
  - Information retrieval
  - Navigation systems
  - Advertising
  - Security
- Generating images and video from text
  - Computer animation

Lecturers

Simon

John

Mehdi

Course webpage and materials

UNIVERSITY OF GOTHENBURG

https://www.dobnik.net/simon/events/apl-esslli-19/

apl-esslli@dobnik.net

CLASP centre for linguistic theory and studies in probability

Tutorials and practical exercises

UNIVERSITY OF GOTHENBURG

- Google Colab
- Computer with GPU, Python3 with jupyter-notebook, pillow, matplotlib, tensorflow, keras

Language, Action, and Perception

CLASP centre for linguistic theory and studies in probability

How do we do it?

1. 

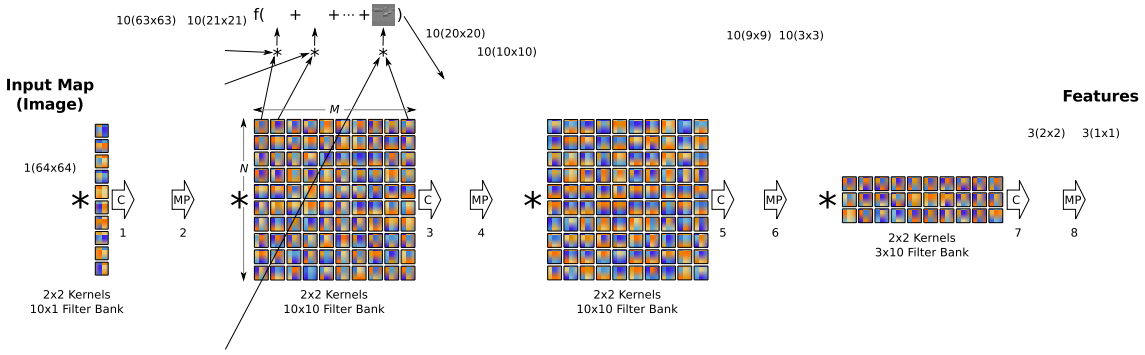
How do we do it?

1. 
2. *The newspaper is on the table*

How do we do it?

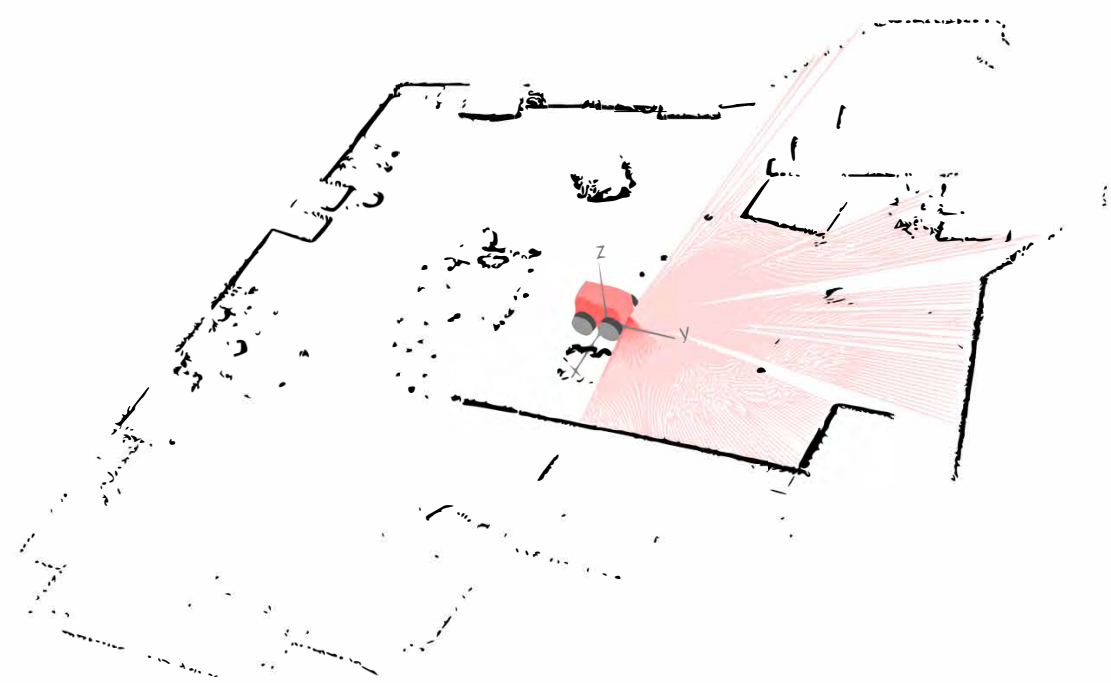
1. 
2. *The newspaper is on the table*
3. $\forall x \forall y[\text{supports}(y,x) \wedge \text{contiguous}(\text{surface}(x), \text{surface}(y)) \rightarrow \text{on}_1(x,y)]$

How do we do it?

1. 
2. *The newspaper is on the table*
3. $\forall x \forall y[\text{supports}(y,x) \wedge \text{contiguous}(\text{surface}(x), \text{surface}(y)) \rightarrow \text{on}_1(x,y)]$
4. 

Picture from (Koutník et al., 2014)

How do we do it?

1. 

2. *The newspaper is on the table*
3. ...

Some key questions

UNIVERSITY OF GOTHENBURG

- How does natural language interact with the physical world through action and perception?
  - How a situated agent can make sense of the world/assign meaning in which it is located?
  - How a situated agent can make sense of the conversation with other situated agents?
- How to mediate between perceptual sensory data (real numbers) and symbolic representations of language?
- How to deal with constantly changing world - learn from experience?

What do we need?

Here are some examples from the Flickr8k corpus (Rashtchian et al., 2010). Each image is followed by five descriptions. The descriptions were made by human annotators using crowd-sourcing with Amazon Mechanical Turk, one description per person per image.

The spatial relations that I would like you to focus on are highlighted. Think about the problems we need to solve to connect words (describing spatial relations) with images.

166321294

- A man is riding on a red motorcycle.
- A motorcycle driver dressed in orange gear swerves to the right.
- A motorcyclist on a red speed bike leans into a sharp turn.
- Motorcyclist crouches low as he rounds a turn.
- This person is on a red motorcycle.

2101128963

- A baseball is recoiling from an action taken on a treated field watched by others.
- A baseball player on a playing field springs into action.
- A baseball player standing on the mound.
- A Philadelphia Phillie pitcher on the pitchers mound with his left leg up behind him.
- A pitcher in a red and white uniform in a baseball game has just thrown the ball.

3030824089

- A big black and brown dog plays outdoors.
- A black and tan dog leaps over the green grass.
- A brown and black dog runs on the grass outdoors in front of a sidewalk.
- A dog runs.
- A German shepherd jumps left on patchy grass.

Simon's summary

- Theoretical background
  - How words associate with pixels
  - Objects (rules of physics, interaction between objects, what is their function)
  - Geometry and relations
  - Perspective
  - What the image is about - attention?

Simon's summary

- Theoretical background
  - How words associate with pixels
  - Objects (rules of physics, interaction between objects, what is their function)
  - Geometry and relations
  - Perspective
  - What the image is about - attention?
- Representations and algorithms
  - Computational models of language and perception
  - Representations and information fusion
  - Machine learning from examples
  - Integration of background knowledge

Simon's summary

- Theoretical background
  - How words associate with pixels
  - Objects (rules of physics, interaction between objects, what is their function)
  - Geometry and relations
  - Perspective
  - What the image is about - attention?
- Representations and algorithms
  - Computational models of language and perception
  - Representations and information fusion
  - Machine learning from examples
  - Integration of background knowledge
- Applications
  - Generating image descriptions (NLG)
  - Visual question answering (NLU and NLG)

Representations of meaning

#1 Model-theoretic meaning

UNIVERSITY OF GOTHENBURG

- Model: state of affairs external to any agent
  $\{ \mathsf{l}, \mathsf{a}, \mathsf{g}, \mathsf{s}, \langle \mathsf{a} \rangle, \langle \mathsf{l} \rangle \langle \mathsf{a}, \mathsf{g} \rangle, \langle \mathsf{s}, \mathsf{l} \rangle, \ldots \}$
- There may be sets of related but slightly different models: possible worlds
- Linguistic expressions $\rightsquigarrow$ expressions of a formal language
- Expressions :: truth conditions
- Interpretation function
- Compositionality
- Typed Lambda Calculus: types e and t and function types

(Montague, 1973; Dowty et al., 1981; Blackburn and Bos, 2005)

Questions not answered

UNIVERSITY OF GOTHENBURG

- How are models built?
- How are assignments determined?
- What happens if the world/the usage of a word changes?
- Is meaning really external - interaction between agents?
- How do we build all possible models/worlds (before running out of memory)?

#2 Distributional meaning

Distributional hypothesis of lexical meaning

- The meaning of a word is the set of contexts in which it occurs
- Important aspects of the meaning of a word are a function of (can be approximated by) the set of contexts in which it occurs in texts

#2 Distributional meaning

UNIVERSITY OF GOTHENBURG

Distributional hypothesis of lexical meaning

- The meaning of a word is the set of contexts in which it occurs
- Important aspects of the meaning of a word are a function of (can be approximated by) the set of contexts in which it occurs in texts

1. He filled the wampimuk, passed it around and we all drank some.
2. We found a little, hairy wampimuk sleeping behind the tree.

#2 Distributional meaning, II

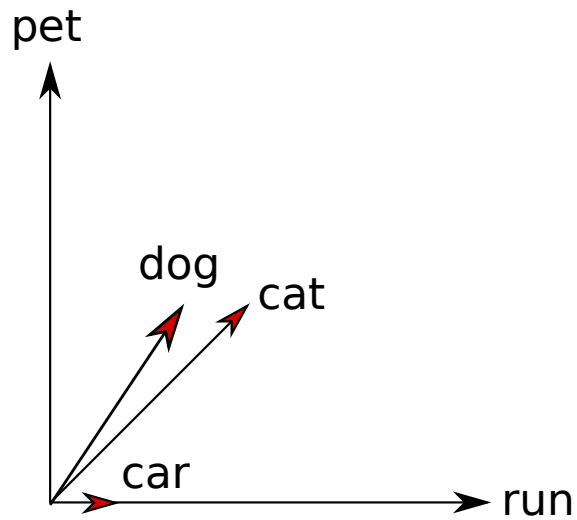
UNIVERSITY OF GOTHENBURG

- Collect a corpus of text
- Represent the meaning of words as context-word vectors representing the distribution of a word

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
...						
car	0	0	1	3	0	0

#2 Distributional meaning, III

- Use geometric methods on vectors to determine distance in space defined by distributional vectors (cosine similarity)



- Connect distributional tensors of word contexts with types/categories to ensure compositionality

(Turney et al., 2010; Clark, 2015; Mitchell and Lapata, 2010; Coecke et al., 2010)

#2 Distributional meaning, IV

- Use ML to learn contextual generalisations: neural language models

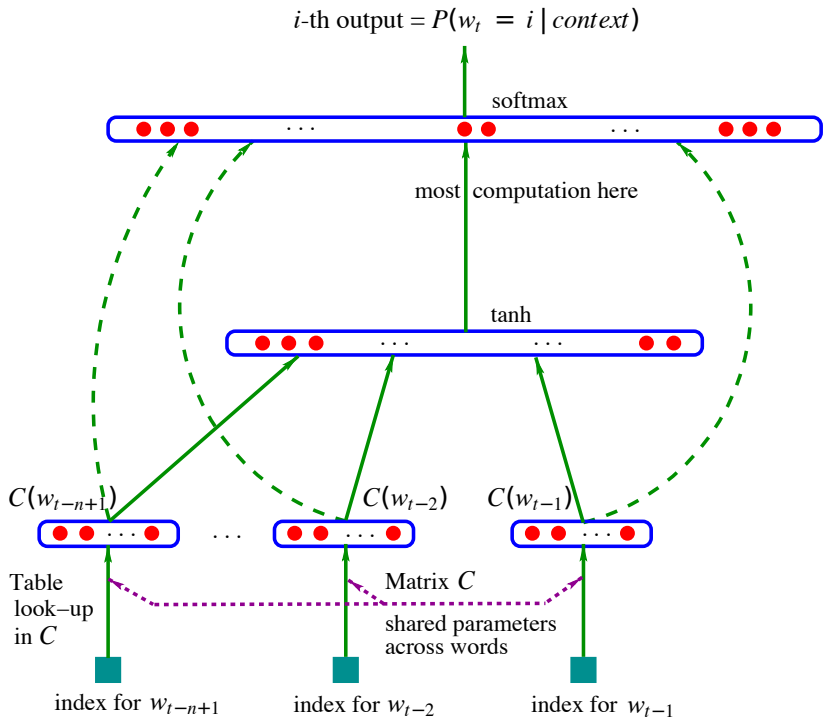


Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i

(Bengio et al., 2003; Peters et al., 2018; Devlin et al., 2018)

Questions not answered

UNIVERSITY OF GOTHENBURG

- Tell us a lot about the world
  situations $\rightsquigarrow$ descriptions of situations $\rightsquigarrow$ distributional representations
- Disconnected from the world
  cf. Chinese room argument (Searle, 1980)
- How can we evaluate linguistic expressions as being true or false?

Questions not answered

UNIVERSITY OF GOTHENBURG

- Tell us a lot about the world
  situations $\rightsquigarrow$ descriptions of situations $\rightsquigarrow$ distributional representations
- Disconnected from the world
  cf. Chinese room argument (Searle, 1980)
- How can we evaluate linguistic expressions as being true or false?
  - The sun rises in the East.
  - Riga lies on the Gulf of Riga at the mouth of the Daugava river where it meets the Baltic Sea.

Questions not answered

UNIVERSITY OF GOTHENBURG

- Tell us a lot about the world
  situations $\rightsquigarrow$ descriptions of situations $\rightsquigarrow$ distributional representations
- Disconnected from the world
  cf. Chinese room argument (Searle, 1980)
- How can we evaluate linguistic expressions as being true or false?
  - The sun rises in the East.
  - Riga lies on the Gulf of Riga at the mouth of the Daugava river where it meets the Baltic Sea.
  - The chair is to the left of the table.
  - The chair is to the right of the table.
- Compositionality?

#3: Grounded and embodied meaning

UNIVERSITY OF GOTHENBURG

- Humans "can (1) discriminate, (2) manipulate (3) identify and (4) describe the objects, events and states of affairs in the world they live in, and they can also (5) produce descriptions and (6) respond to descriptions of those objects, events and states of affairs." (Harnad, 1990, p.341)
- Embodied mind (Maurice Merleau-Ponty and George Lakoff)

#3: Grounded and embodied meaning

UNIVERSITY OF GOTHENBURG

- Humans "can (1) discriminate, (2) manipulate (3) identify and (4) describe the objects, events and states of affairs in the world they live in, and they can also (5) produce descriptions and (6) respond to descriptions of those objects, events and states of affairs." (Harnad, 1990, p.341)
- Embodied mind (Maurice Merleau-Ponty and George Lakoff)
- Language/cognition vs sensory representations

Sensory readings	Human language
Continuous measures	Discrete categories
Accurate reference	Underspecified reference
Mathematical representations	Cognitive representations

#3: Grounded and embodied meaning, II

(Harnad, 1990)

- Types of representations:
  - Iconic representations
  - Categorical representations
  - Higher level symbolic representations: compositional structure
- Representations are connected by learning through classification

#3: Grounded and embodied meaning, III

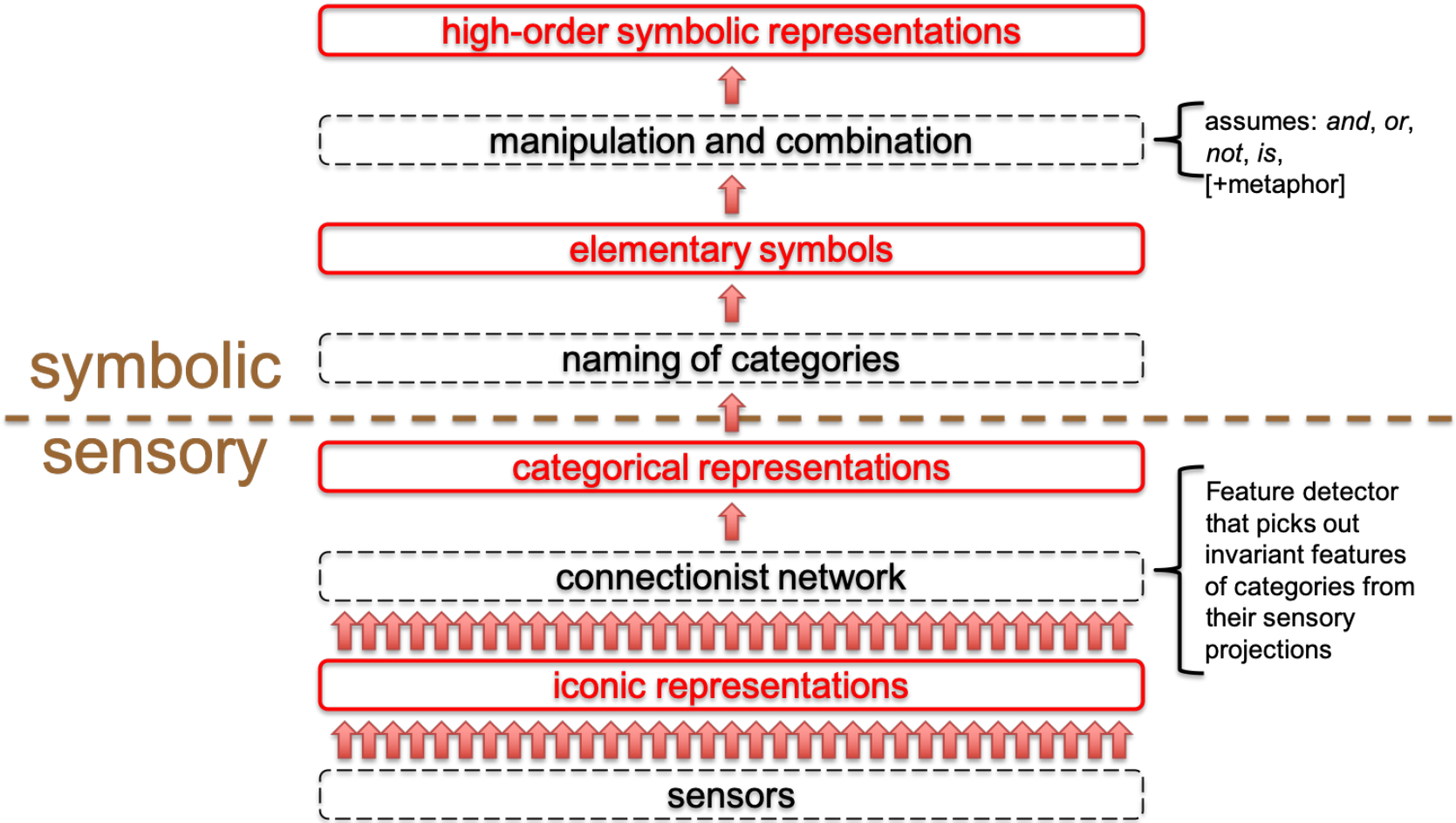
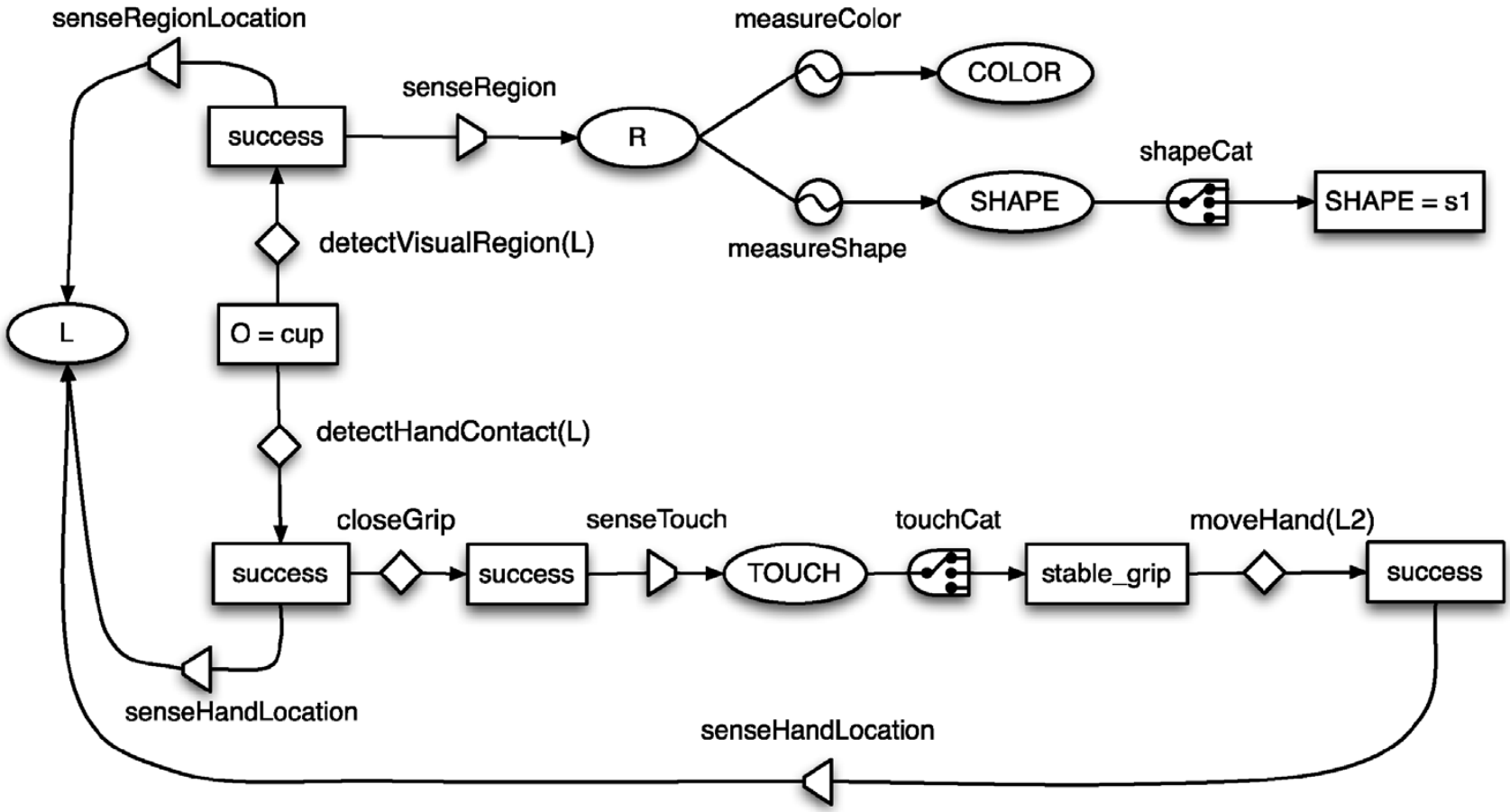


Image from Kelleher (2010)

#3: Grounded and embodied meaning, IV



(Roy, 2005)

Embodiment

- Meaning is internal to the agent
- Agents with different bodies (sensors and actuators) perceive an interact with the world differently.
- Consequently, they also structure the world differently: the representations they learn will be different
- Is human-robot communication possible at all?

Situatedness and interaction

UNIVERSITY OF GOTHENBURG

- Human and robot are situated in the same environment which imposes identical constraints on both kinds of representations.
- They can also interact with each other: see each other, jointly attend to each other and refer to the same situations.
- Grounded language models must be continuously adapted
- Perhaps the fact that they may internally operate with different representations is not that important.

Questions not answered

UNIVERSITY OF GOTHENBURG

- Compositionality is a property of symbolic systems (?)
- Different words are grounded in perception and action to a different degree
- Some aspects of meaning are not grounded
- Different representations of meaning are complementary in strengths and weaknesses

Coming up next in the coming days

1. ...
2. Language and space
3. Generating and interpreting grounded language
4. Referring to what matters (attention)
5. Learning language with robots

References I

Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*(6), 1137–1155.

Blackburn, P. and J. Bos (2005). *Representation and inference for natural language. A first course in computational semantics*. CSLI Publications.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantics — second edition*, Chapter 16, pp. 493–522. Wiley – Blackwell.

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv arXiv:1003.4394 [cs.CL]*, 1–34.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv arXiv:1810.04805 [cs.CL]*, 1–14.

Dowty, D. R., R. E. Wall, and S. Peters (1981). *Introduction to Montague semantics*. Dordrecht, Holland: D. Reidel Pub. Co.

References II

Harnad, S. (1990, June). The symbol grounding problem. *Physica D 42*(1–3), 335–346.

Kelleher, J. D. (2010). How to preposition a robot: A case-study in symbol grounding. Presentation at the University College Dublin Research Seminars.

Koutník, J., J. Schmidhuber, and F. Gomez (2014). Evolving deep unsupervised convolutional networks for vision-based reinforcement learning. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pp. 541–548. ACM.

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8), 1388–1429.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In J. Hintikka, J. Moravcsik, and P. Suppes (Eds.), *Approaches to Natural Language*, pp. 221–242. Dordrecht. reprinted in Thomason, 1974.

References III

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 2227–2237. Association for Computational Linguistics.

Rashtchian, C., P. Young, M. Hodosh, and J. Hockenmaier (2010, 6 June). Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

Roy, D. (2005, September). Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence 167*(1-2), 170–205.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417–424.

References IV

Turney, P. D., P. Pantel, et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*(1), 141–188.