Understanding Spoken Language Understanding

from big data to real-world context

Prof. Roger K. Moore

Chair of Spoken Language Processing
Dept. Computer Science, University of Sheffield
(Visiting Prof., Dept. Phonetics, University College London)
(Visiting Prof., Bristol Robotics Lab.)

Spoken Language Understanding

The 'State-of-the-Art'

- There is steady year-on-year progress
- Improvement are coming from:
  - corpus-driven **statistical modelling** approaches
  - public benchmark testing
  - increase in available computer power
- Progress has *not* come about as a result of deep insights into human spoken language
- Spoken language technology is
  - **fragile** (*in 'real' conditions*)
  - **expensive** (*to port to new applications / languages*)
- ASR performance is reaching an *asymptote* well short of human abilities
  - 25% word error rate on conversational speech

Inhibits creativity?

Over-reliance on data?

The 'State-of-th

- ere is steady year-on-ye ogress
- Impr ement are coming om:
  - corpus-driven **statistical modelling** approaches
  - public benchmark testing
  - increase in available computer power
- Progress has *not* come abou as a result of deep insights into human spok language
- Spoken language technology is
  - **fragile** (*in 'real' conditions*)
  - **expensive** (*to port to new applicatio*
- ASR performance is reaching an short of human abilities
  - 25% word error rate on conversational speech

Too easy to 'tweak' algorithms?

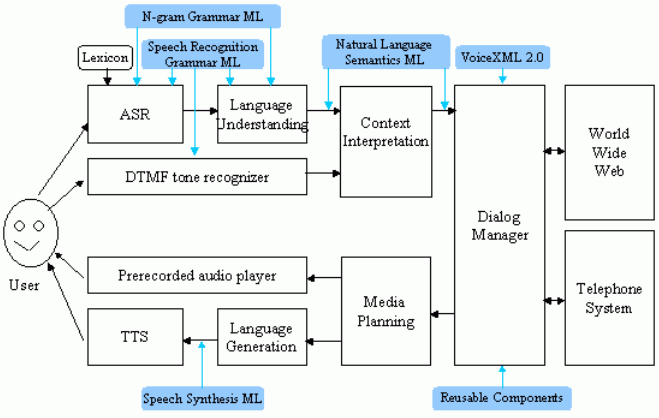
The 'State-of-the-Art'

- There is steady year-on-year progress
- Improvement are coming from:
  - corpus-driven **statistical models** approaches
  - public benchmark testing
  - increase in available computer power
- Progress has *not* come about as a result of deep insights into human spoken language
- Spoken language technology is
  - **fragile** (*in 'real' conditions*)
  - **expensive** (*to port to new applications / languages*)
- ASR performance is reaching an *asymptote* well short of human abilities
  - 25% word error rate on conversational speech

New opportunities?

Missed opportunities?

'Traditional' Architecture



Introduction and Overview of W3C Speech Interface Framework, http://www.w3.org/TR/voice-intro/

Insights from Living Systems

Perceptual Control

Emotion

Illusion of Invariance

Particulate structures

Enactivism

Imitation

Emulation

Autopoiesis

Memory

Theory of Mind

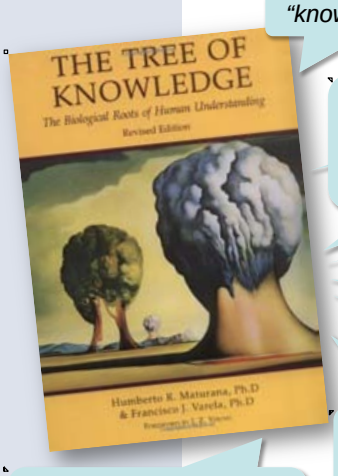
Prediction

Mirror Neurons

Sensorimotor Overlap

The University Of Sheffield.

SPandH



"knowing is doing"

"human linguistic behaviours are ... in a domain of reciprocal ontogenetic structural coupling"

"everything we do is a structural dance in the choreography of coexistence"

"'enactive' … what is known is brought forth"

"the phenomenon of communication depends on not what is transmitted, but on what happens to the person who receives it"

"cognition is based on the organism as a unity and on the operational closure of its nervous system"

"living beings are autonomous unities"

"we maintain an ongoing descriptive recursion which we call 'I'"

Maturana, H. R., & Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications.

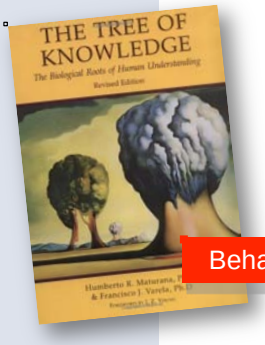
The University Of Sheffield.

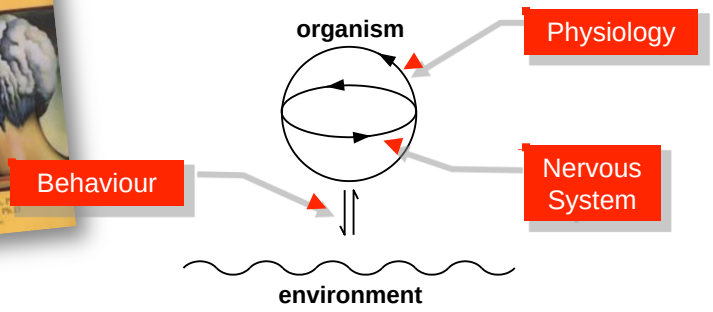
SPandH

Enactivism

A 'cognitive unity'

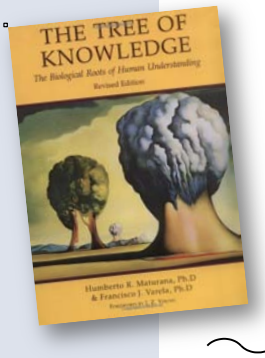
(*self-regulating self-generating closure*)

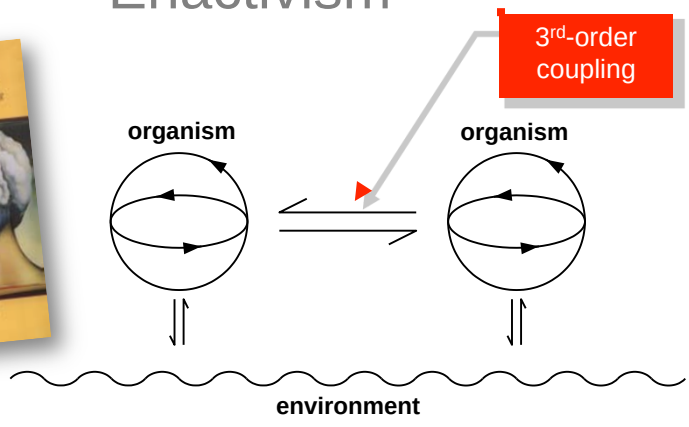




Maturana, H. R., & Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications.

Enactivism





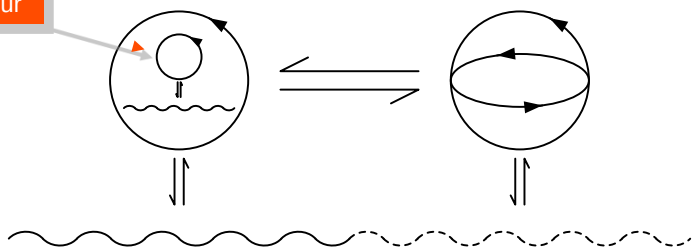
Maturana, H. R., & Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications.

Emulation

Modelling the 'surface behaviour' of another agent

Predicting other's external behaviour

Similar to Dennett's '*physical stance*'



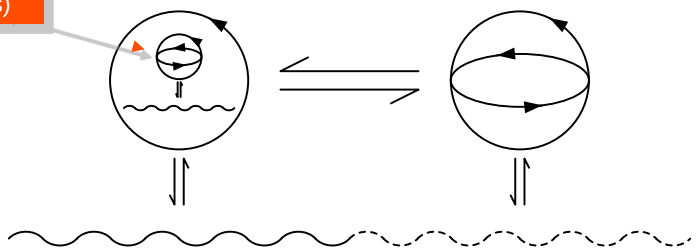
Moore, R. K. (2012). Extending Maturana and Varela's symbolic representation of autopoiesis to create a rich visual language for envisioning a wide range of enactive systems with different degrees of complexity, *Foundations of Enactive Cognitive Science*. Cumberland Lodge, Great Park of Windsor.

Empathy

Modelling the 'internal states' of another agent

Predicting other's affective state(s)

Similar to Dennett's '*design stance*'



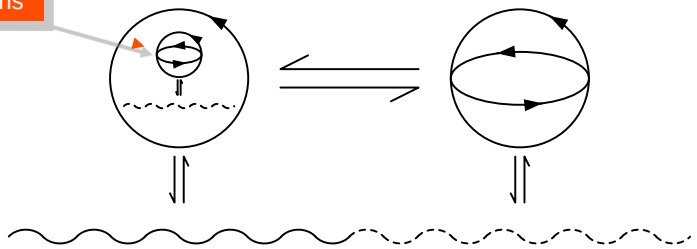
Moore, R. K. (2012). Extending Maturana and Varela's symbolic representation of autopoiesis to create a rich visual language for envisioning a wide range of enactive systems with different degrees of complexity, *Foundations of Enactive Cognitive Science*. Cumberland Lodge, Great Park of Windsor.

Theory of Mind

Modelling the '**perspective**' of another agent

Predicting other's beliefs & intentions

Similar to Dennett's '*intentional stance*'

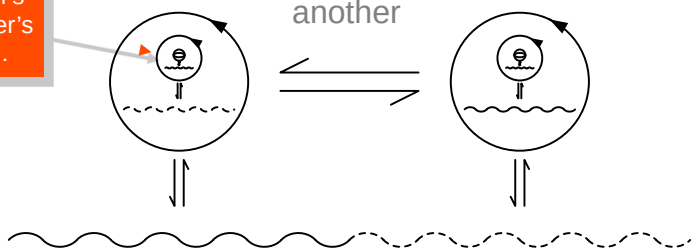


Moore, R. K. (2012). Extending Maturana and Varela's symbolic representation of autopoiesis to create a rich visual language for envisioning a wide range of enactive systems with different degrees of complexity, *Foundations of Enactive Cognitive Science*. Cumberland Lodge, Great Park of Windsor.

Languaging

Modelling the '**recursive particulate coupling**' between one agent and another

Predicting other's knowledge of self's knowledge of other's knowledge of …



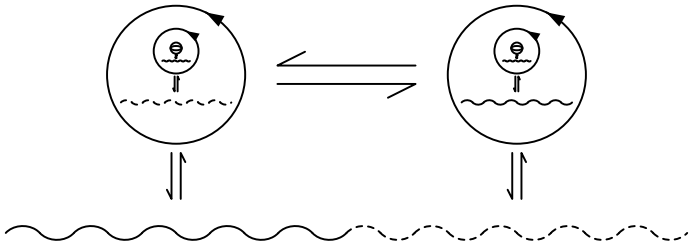
Moore, R. K. (2012). Extending Maturana and Varela's symbolic representation of autopoiesis to create a rich visual language for envisioning a wide range of enactive systems with different degrees of complexity, *Foundations of Enactive Cognitive Science*. Cumberland Lodge, Great Park of Windsor.

Languaging

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04), 329–347.

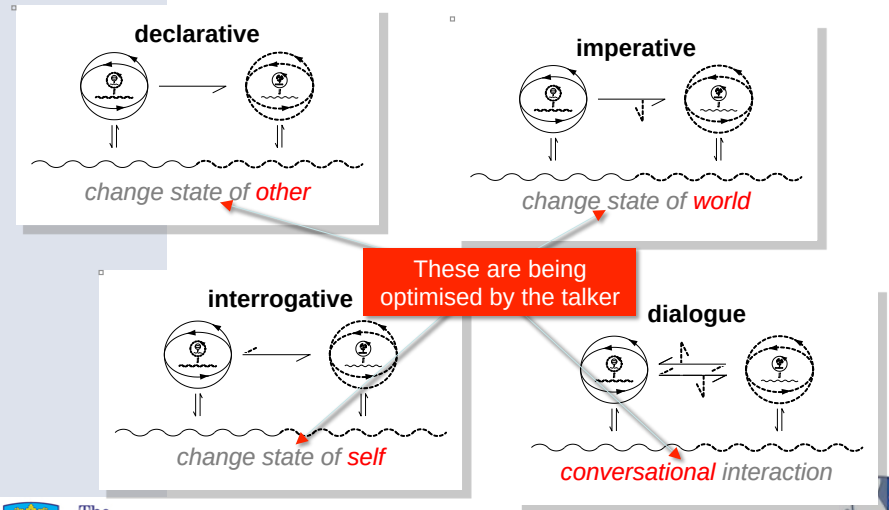
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204.

If a sender knows that the receiver has a predictive model of the talker, …



… then only the difference between the intentions and the predictions need to be sent (*thereby minimising the entropy*)

Languaging



Cognitive Neuroscience

"recognising the intentions of others is based … on the same mechanisms underlying the formation of one's own motor intention"

Frith, C. D., & Lau, H. C. (2006). The problem of introspection. ***Consciousness and Cognition***, 15, 761-764.

Forward/Generative Model (*predictor*)

Model of '**self**'

'Prior' Probability

$$\Pr(intentions \mid behaviour) = \frac{\Pr(behaviour \mid intentions)\Pr(intentions)}{\Pr(behaviour)}$$

'Posterior' Probability

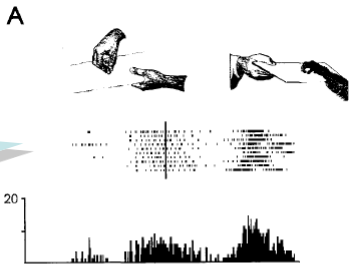
Interpretation of '**other**'

Possible mechanism … '*Mirror Neurons*'

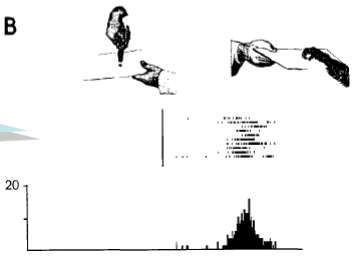
Wilson, M., Knoblich, G., 2005. The case for motor involvement in perceiving conspecifics. ***Psychological Bulletin*** 131 (3), 460-473.

'Mirror' Neurons

A: Neurons fire when the Monkey sees the experimenter grasp the raisin *and* when the Monkey grasps the raisin



B: Neurons do *not* fire when the Monkey sees the experimenter grasp the raisin with a tool



Rizzolatti, G., Fadiga, L, Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. ***Cognitive Brain Research*** 3, 131-141.

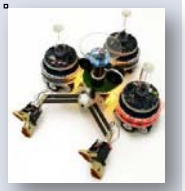
Mirror Mechanisms

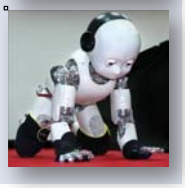
It is hypothesised that projecting '**self**' onto '**other**' facilitates …

- interpretation of intentions
- action understanding
- imitation
- learning (*by imitation*)
- empathy
- Theory of Mind (ToM)
- gestural communication
- evolution of speech & language

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27, 169-192.

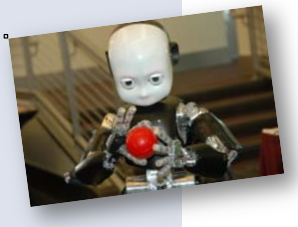
Insights from Cognitive Systems





Embodiment
Robotics
Grounding
Action Selection
Situatedness
Genetic Algorithms
DNNs
Human-Robot Interaction
Sensorimotor coordination
Control Theory
Simulation
Uncanny Valley
Action Understanding

Embodied Cognition

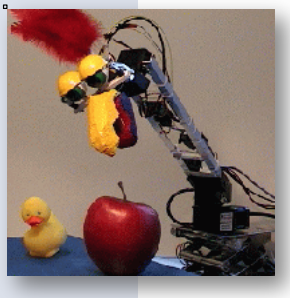


- Embodiment
  - multimodal sensorimotor experience
  - real-world constraints (*laws of physics*)
  - structural coupling
- Situatedness
  - evolution of events
  - interactional history (*memory*)
  - imagination (*prediction*)

Physical Context

Temporal Context

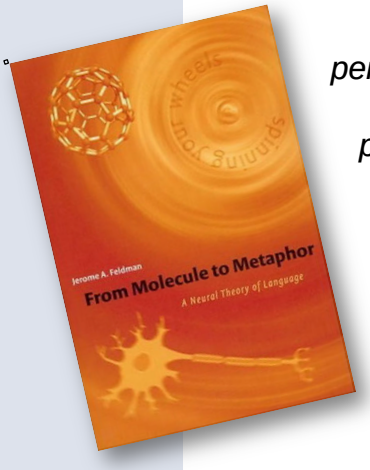
Language Grounding



- Symbol grounding through models of situated/embodied action/perception
- Abstraction facilitated by metaphor
- Decoding performed by simulation

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

Metaphor



"Understanding language about perceiving and moving involves much of the same neural circuitry as do perceiving and moving themselves."

"The embodied neural approach to language suggests that the complex neural circuitry that supports grasping is the core meaning of the word."

Feldman, J. A. (2008). *From Molecules to Metaphor: A Neural Theory of Language*. Bradford Books.

Simulation

- Simulation facilitates …
  - the prediction of future events
  - the explanation of observed events
  - the imagination of novel events
  - the optimal influence of future events
- Simulation implies …
  - generative models of spoken language production (*derived from models of movement and action*)
  - revisiting the motor theory of speech (*i.e. model of 'self' recruited to model 'other'*)

Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. *Speech Communication*, 49(5), 418–435.

Feedback

- The structural coupling of an agent with its environment (*including other agents*) implies feedback
- Feedback facilitates …
  - the maintenance of stability
  - the comparison of achievements against intentions, which …
    - creates affective states, which …
    - drives internal/external behaviour (*including adaptation and learning*)

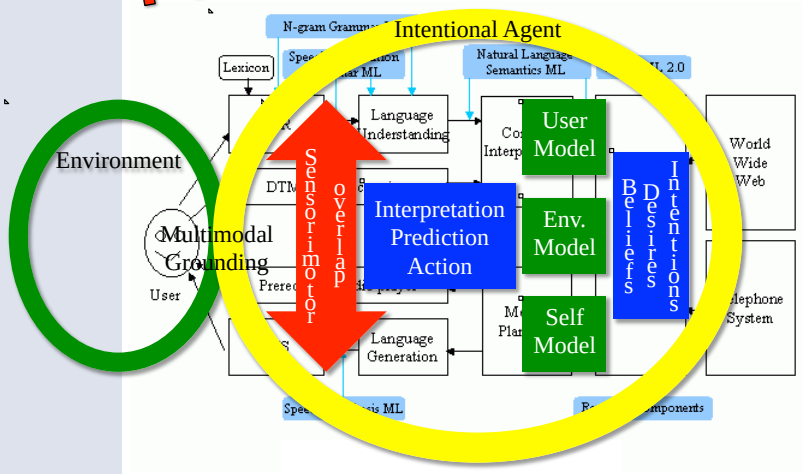
Summary

- What are we doing right?
  - using DNNs (*for non-linear transformations*)
  - using stochastic modelling (*to capture variability*)
  - treating recognition as 'search' (*through a generative data structure*)

- What are we doing wrong?
  - off-line training using ecologically unrealistic amounts of unrepresentative data
  - ignoring speaker-listener-environment coupling
  - ignoring communicative intent
  - treating understanding as 'search' (*through an associative data structure*)
  - treating recognition/understanding independently from generation/synthesis

'Traditional' Architecture

Future



Conclusion

- Lots of new ideas from outside the field, *e.g.* …
  - sensorimotor overlap removes the independence between traditional components
  - focus attention on generative models of spoken language production (*derived from models of movement and action*)
  - replace traditional learning paradigms with on-line interactive skill acquisition in real-world situations and environments
- Challenges …
  - model parameter estimation in a dynamic interactive context
  - evaluation paradigms and metrics
- Progress will almost certainly require an interdisciplinary approach


Thank You
http://www.dcs.shef.ac.uk/~roger