Memory and Attention and Situated Dialog

John D. Kelleher

Language, Action, and Perception
31 European Summer School in Logic,
Language and Information
12th – 16th August 2019, Riga Latvia

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN

Outline

- Situated Dialog
- Reference in Situated Dialog
- Cognitive Theories of Memory
- Grounding Language in Vision
- Visual Attention
- Computational Models of Multimodal Working Memory for Dialog
  - A Local/Episodic Architecture
  - A Global/Monolithic Architecture
- Summary

Situated Dialog

Situated Dialog

- Situated language is spoken from a particular point of view within a shared perceptual context (Byron, 2003)

Situated Dialog

- The history of computational models of situated dialog can be traced back to systems in the 1970's such as SHRDLU Winograd (1973).

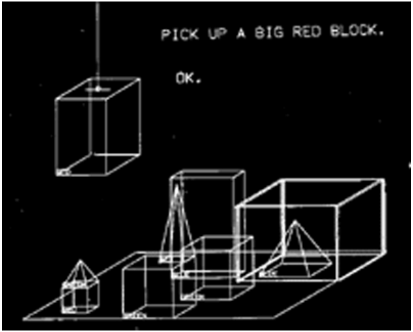


Figure: SHRDLU robot world

See Kelleher and Dobnik (2019) and references therein for a review of relevant literature on situated dialog systems.
SHRLDU image sourced from: https://pl.wikipedia.org/wiki/SHRDLU

Situated Dialog

- A commonality across many of these systems is that they have a primary focus on grounding (in the sense of Harnad (1990) rather than Clark et al. (1991)), the references within a single utterance against the current perceptual context.
- Consequently a key challenge that these systems address is reference resolution

Reference in Situated Dialog

Reference in Situated Dialog

Referring expression can take a variety of surface forms, including:

- definite descriptions: *the red chair*
- indefinites: *a chair*
- pronouns: *it*
- demonstratives: *that*

Reference in Situated Dialog

- The form of referring expression used by a speaker signals their belief with respect to the status the referent occupies within the hearer's set of beliefs
- For example, a pronominal reference signals that the intended referent has a high degree of salience within the hearer's current mental model of the discourse context.

Reference in Situated Dialog

- Mutual knowledge: the set of things that are taken as shared knowledge by interlocutors, and hence are available as referents within the discourse (McCawley, 1993).
- An interlocutor may consider an entity to be in the mutual knowledge set if:
  - it is part of the assumed cultural or biographical knowledge they share with their dialog partner
  - it is in the shared perception of the situation the dialog occurs within.

Reference in Situated Dialog

- The term discourse context (DC) is often used in linguistically focused research on dialog to describe the set of entities available for reference due to the fact that they have previously been mentioned in the dialog

Reference in Situated Dialog

- Mutual Knowledge v. Discourse Context
  - mutual knowledge: the set of entities that are available for reference but which have not been mentioned previously in the discourse
  - discourse context: a record of the entities that have been mentioned previously
- This distinction opens up the possibility that the internal structure of these two components may be quite distinct, we will return to this question later.

Reference in Situated Dialog

The process of resolving a referring expression can be characterized as follows:

1. a referring expression in an utterance introduces a representation into the semantics of that utterance
2. this representation must be bound to an entity in the mutual knowledge set or in the discourse context for the utterance to be resolved.

Reference in Situated Dialog

We can distinguish three types of referring expressions based on the information source they draw their referent from (as opposed to their surface form), namely:

1. evoking,
2. exophoric,
3. anaphoric

Reference in Situated Dialog

- An evoking reference refers to an entity that is known to the interpreter through their conceptual knowledge but which has not previously been mentioned in the dialog.
- The referent of an evoking reference is found in the mutual knowledge set
- The process of resolving this reference introduces a representation of the referent into the discourse context.

Reference in Situated Dialog

- An exophoric reference denotes an entity that is known to the interpreter through their perception of the situation of the dialog but which has not previously been mentioned in the dialog.
- Similar to an evoking reference, the process of resolving an exophoric reference introduces a representation of the referent into the discourse context.

Reference in Situated Dialog

- An anaphoric reference refers to an entity that has already been introduced mentioned in the dialog and hence a representation of its referent is already in the discourse context.

Reference in Situated Dialog

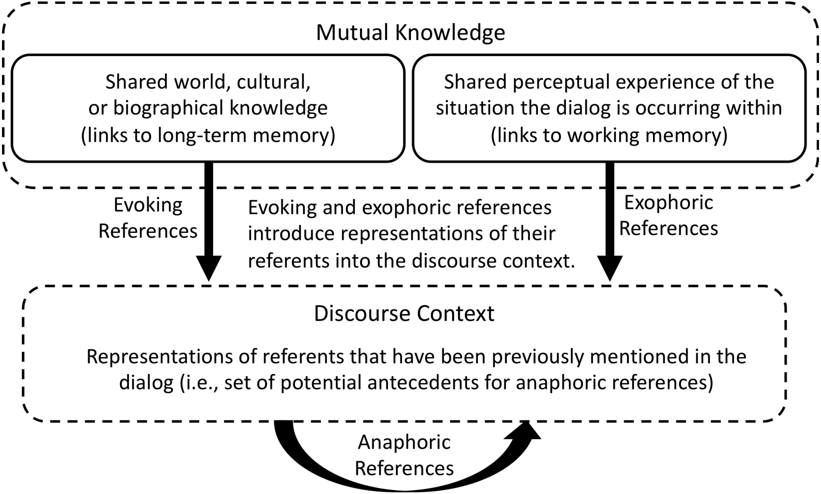

Mutual Knowledge
Shared world, cultural, or biographical knowledge (links to long-term memory)
Shared perceptual experience of the situation the dialog is occurring within (links to working memory)
Evoking References
Evoking and exophoric references introduce representations of their referents into the discourse context.
Exophoric References
Discourse Context
Representations of referents that have been previously mentioned in the dialog (i.e., set of potential antecedents for anaphoric references)
Anaphoric References

Figure: The relationship between mutual knowledge, the discourse context, and evoking, exophoric, and anaphoric references.

Reference in Situated Dialog

- All of these form of reference draw upon human memory.
- Mutual knowledge and the maintenance of a discourse context are both stored in memory.

Cognitive Theories of Memory

Cognitive Theories of Memory

Cognitive psychology[1] distinguishes between a number of different types of memory including:

- sensory memory which persists for several hundred milliseconds and is modal specific
- working memory which persists for up thirty seconds and has limited capacity
- long-term memory which persists from thirty minutes to the end of an person's lifetime, and has potentially unlimited capacity.

The Atkinson and Shiffrin (1968) model describes how these different types of memory interact.

[1] See, for example Eysenck and Keane (2013).

Cognitive Theories of Memory

Sensory Memory

- External inputs are initially stored in modality specific sensory memory buffers.
- There is an attentional filter between these sensory specific memories and working memory.
  - Information that is attended to passes through to working memory
  - Unattended information is lost

Cognitive Theories of Memory

Working Memory

- Limited capacity
- Information in the working memory that is frequently rehearsed is transferred to long-term memory and may be retrieved later.
- Information is working memory that is not rehearsed is displaced as new information arrives.

Cognitive Theories of Memory

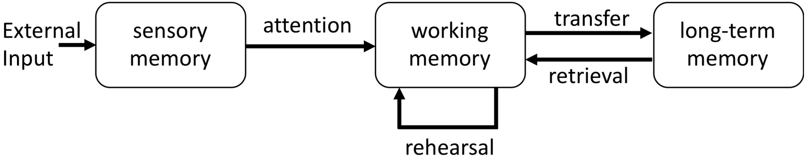


Figure: Multi-store Model of Memory based on Atkinson and Shiffrin (1968)

Cognitive Theories of Memory

- The Atkinson and Shiffrin (1968) Multi-store Model of Memory is the most popular model of memory.
- However, alternative models have been proposed.
- For example, the levels-of-processing (Craik and Lockhart, 1972) is another well-known model that proposes:
  1. there is no compartmental structure to memory (i.e. there is no distinction between different types of memory
  2. and the ability to remember or recall something is dependent on the depth of processing measured on a continuous scale from ranging from 'shallow' (perceptual) to 'deep' (semantic)

Cognitive Theories of Memory

Intermediate-term memory (ITM)

- is a stage of memory distinct from sensory memory, working memory/short-term memory, and long-term memory.
- it persists for about two to three hours.
- it declines completely before the onset of long-term memory

Unlike short-term memory and working memory, intermediate-term memory requires changes in translation to occur in order to function. While ITM requires only changes in translation, induction of long-term memory requires changes in transcription as well. Translation and transcription are concepts from microbiology and genetics.

Cognitive Theories of Memory

- Evoking references draw on long-term memory
- Exophoric references draw on working memory
- It is also reasonable that the discourse context model should be considered a part of working memory/intermediate-term memory

→ Working memory is at crux of handing anaphoric and exophoric references.

Cognitive Theories of Memory

Baddeley (2002) model of working memory has 4 major systems:

- central executive: modality independent, supervises the integration of information, directs attention, coordinates other systems
- phonological loop holds speech based information and can maintain this information over short periods by continuous rehearsal
- visual-spatial sketchpad stores visual and spatial information and can construct visual images and mental maps
- episodic buffer
  - a limited capacity buffer
  - temporarily stores and integrates information from other modules (phonological loop, the visuo-spatial sketchpad, smell, taste, and so on)
  - integrates disparate encodings into a unitary representation of chronologically ordered episodes.

Cognitive Theories of Memory

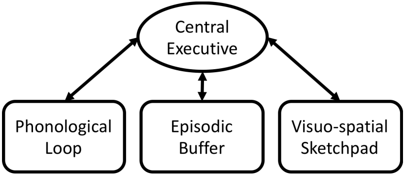


Figure: Baddeley's Model of Working Memory

Grounding Language in Vision

Grounding Language in Vision

- Grosz (1977) highlighted that attention processes can affect how references are resolved during a dialog.
    - if an object is in the mutual focus of attention it can be denoted by means of a definite description even though other entities fulfilling the description are present in the mutual context set.

Grounding Language in Vision

- Grosz and Sidner (1986) extended this work and developed a focus stack model of global discourse attentional state.
- Other models of global discourse structure and processing have since been proposed, for example Hobbs (1985); Mann and Thompson (1987); Kempson (1988); Kempson et al. (2000); Asher and Lascarides (2003); Kamp et al. (2011).

Grounding Language in Vision

- Whichever model of global discourse structure is assumed the question of how the focus of attention and reference interact within a local discourse context must also be addressed
- A number of approaches to this question have been proposed, for example Alshawi (1987), Hajicová (1993), Lappin and Leass (1994), and Grosz et al. (1995).
- However, none of these models explicitly accommodate multimodal contexts

Grounding Language in Vision

Examples of work on reference in mulitmodal contexts:

- Harnad (1990) addresses the question of grounding language in perception.
- Coradeschi and Saffiotti (2003) has addressed this in terms of the symbol anchoring framework
- Roy (2005) has proposed semiotic schemas
- Kruijff et al. (2006) proposed an ontology-based mediation between content in different modalities

Generally, these works focus on exophoric references but assume that the referent is still perceptually available.

Grounding Language in Vision

- An exophoric reference can denote an entity that is not perceptually available at the time of the reference.

Example

- Two people are in a car that is driving along the road.
- passenger: *did you see the cyclist with dog at the traffic lights back there?*
- driver *yes*

- This example highlights the fact that the need for a memory of perception to be maintained to handle these references
- We will refer to these types of exophoric references as references to perceptual memories.

Grounding Language in Vision

- For a system to handle exophoric references to perceptual memories requires the design of a perceptual memory data structure.
- This perceptual memory data structure stores the mutual knowledge information related to the interlocutors shared perceptual experience of the situation
- This perceptual memory data can be understood as part of working memory

Grounding Language in Vision

The design of a perceptual memory data-structure opens up a number of questions, for example:

- should all entities that are perceived be entered into this data structure or is their a filtering process (e.g. an attentional filter)?
- once and entity enters the perceptual memory is it there indefinitely or can it be removed (forgotten)?
- how does the perceptual memory interact with the linguistic discourse history (are they separate)?
- how is the perceptual memory structured, for example, is it episodic or monolithic, does it have a chronological order?
- ...

Visual Attention

Visual Attention

- The human faculty of attention is the “selective aspect of processing” (Kosslyn, 1994, pg. 84)
- Visual Attention regulates the processing of perceived visual stimuli by selecting a region within the visual buffer for detailed processing.

Visual Attention

- In computational systems, a saliency map is used to estimate the regions within an image that receive visual attention

Visual Attention

There are many different and sometimes competing factors that affect the location of the region a perceiver attends to:

- top-down: visual familiarity, intentionality, and so on.
- bottom-up: colour, movement, singleton, and so on.

Visual Attention: Bottom-Up

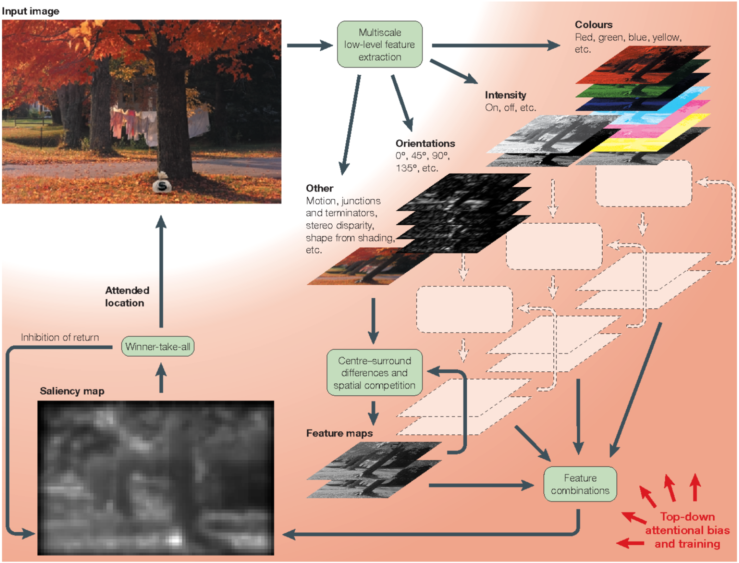

Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

Figure: Figure 1 from Itti and Koch (2001)

Computational Models of Multimodal Working Memory for Dialog

A Local/Episodic Architecture

- The LIVE system Kelleher et al. (2005), is a candidate architecture for this episodic buffer module.
- The LIVE system is designed as a natural language interface to a virtual town, similar in spirit to Winograd's SHRDLU system discussed earlier.
- A distinctive characteristic of the LIVE system, is that the user was able to move around the environment, and the system had a perceptual memory module that enabled the user to refer to off-screen objects that had been seen recently.

A Local/Episodic Architecture

- The LIVE system uses a false colouring visual salience algorithm to process each frame (visual scene) generated as the user moved through the virtual environment Kelleher and van Genabith (2003, 2004), there are 28 such frames generated per second.

Figure: A scene in the LIVE domain.

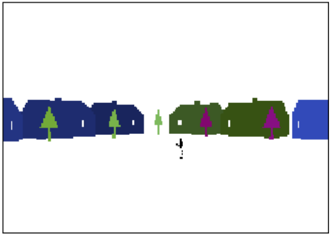


Figure: The false colour rendering of the scene in Figure 6.

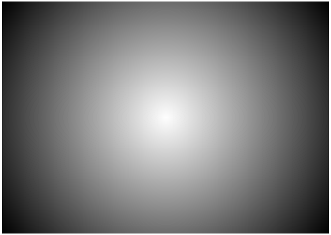


Figure: The weighting assigned to the pixels in the viewport: the darker the pixels the lower the weighting

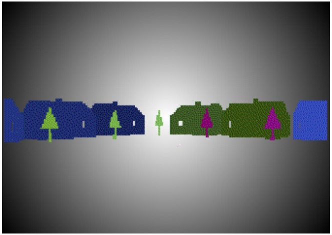


Figure: An overlay of the false colour rendering of Figure 6 on the distribution of pixel weightings.

A Local/Episodic Architecture

- This visual salience algorithm identifies each object instance visible in a frame, and associates a relative visual salience score to each object, based on its size and location within the frame.
- For each frame a list of the visible objects along with their type and colour information and a salience score is created.

A Local/Episodic Architecture

- This frame information is then used to populate a data structure, known as a reference domain
- There is a separate reference domain created for each frame.

A Local/Episodic Architecture

- A reference domain is composed of a number of lists, known as partitions, and the elements of each partition is ordered, in descending order, by their visual salience.
- The function of these partitions is to predict the different ways a user may refer to an object in the scene.
- For example:
  - If trees visible in a frame then the corresponding reference domain would include a tree partition listing all the trees visible ordered by their salience
  - If there are red objects in the scene then there would be a red partition listing all the red objects ordered by colour)

A Local/Episodic Architecture

- The set of potential partitions that could be included in a reference domain is huge: e.g. red houses, or green trees, and other combinations of features.
- The LIVE system limits initial set of partitions to categories that are reasonably likely to be preattentively available, namely, object, type, and colour[2]
- Partitions modelling more complex criteria may be created within a reference domain in response to a linguistic utterances

[2] For a discussion on the question of how we might contextually prime an agent to run perceptual classifiers see Dobnik and Kelleher (2016)

A Local/Episodic Architecture

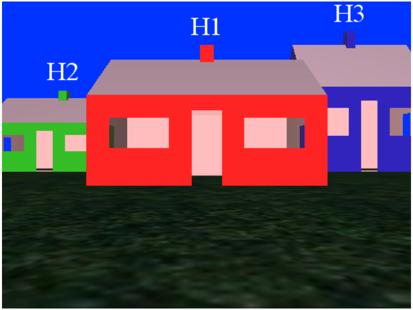
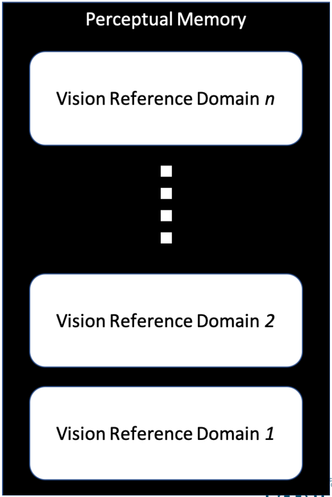


Figure: A frame from the LIVE System.

$$\begin{bmatrix} \text{p1} & \begin{bmatrix} \text{criterion} & \text{`object'} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0; H3,0.2; H2,0.1}\end{bmatrix} \end{bmatrix} \\ \text{p2} & \begin{bmatrix} \text{criterion} & \text{`house'} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0; H3,0.2; H2,0.1}\end{bmatrix} \end{bmatrix} \\ \text{p3} & \begin{bmatrix} \text{criterion} & \text{`red'} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0}\end{bmatrix} \end{bmatrix} \\ \text{p4} & \begin{bmatrix} \text{criterion} & \text{`blue'} \\ \text{elements} & \begin{bmatrix}\text{H3,0.2}\end{bmatrix} \end{bmatrix} \\ \text{p4} & \begin{bmatrix} \text{criterion} & \text{`green'} \\ \text{elements} & \begin{bmatrix}\text{H2,0.1}\end{bmatrix} \end{bmatrix} \end{bmatrix}$$

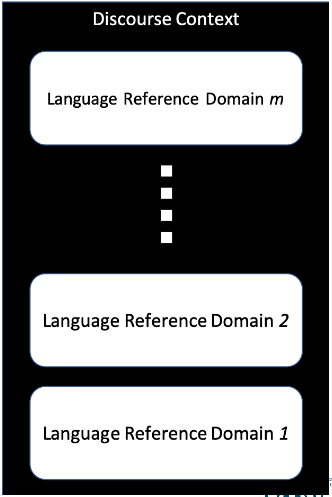
A Local/Episodic Architecture

- The LIVE system stores these reference domains in a chronologically ordered data structure with a capacity to hold 3,000 reference domains
- When the data structure was full the oldest reference domain was deleted to make space for the new reference domain.
- This gives the system a perceptual memory of $\frac{3{,}000}{28} = 108$ seconds.

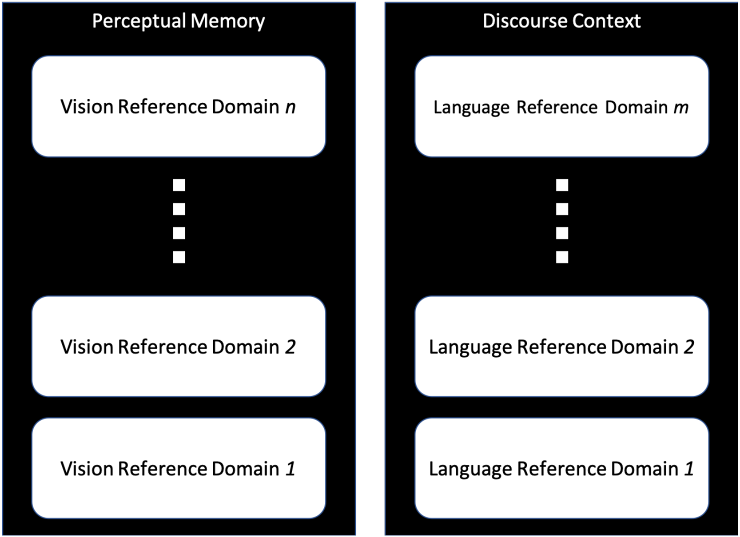

Perceptual Memory
Vision Reference Domain *n*
Vision Reference Domain *2*
Vision Reference Domain *1*

A Local/Episodic Architecture

- The LIVE system also maintained a discourse context model.
- This model is similar in structure to the perceptual memory, it consists of up to 3,000 chronologically ordered reference domain data structures and uses a first-in-first-out policy when the buffer is full.
- New reference domains are added to this discourse context model as a result of resolving a referring expression.


Discourse Context
Language Reference Domain *m*
Language Reference Domain *2*
Language Reference Domain *1*

A Local/Episodic Architecture



Perceptual Memory

Vision Reference Domain *n*

Vision Reference Domain *2*

Vision Reference Domain *1*

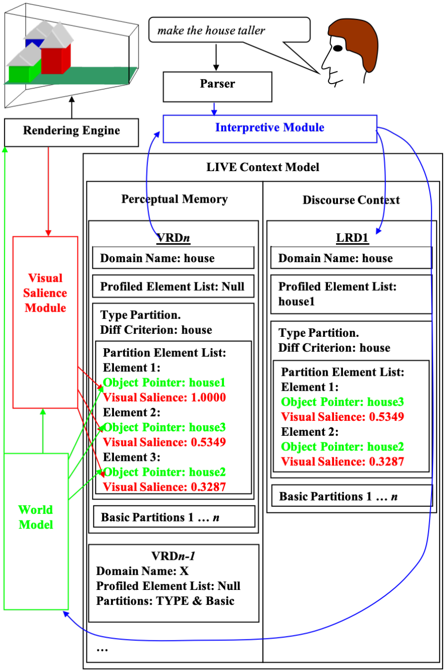
Discourse Context

Language Reference Domain *m*

Language Reference Domain *2*

Language Reference Domain *1*

Figure: LIVE Context Model



make the house taller

Parser

Interpretive Module

Rendering Engine

LIVE Context Model

Perceptual Memory

VRD*n*

Domain Name: house

Profiled Element List: Null

Type Partition.
Diff Criterion: house

Partition Element List:
Element 1:
Object Pointer: house1
Visual Salience: 1.0000
Element 2:
Object Pointer: house3
Visual Salience: 0.5349
Element 3:
Object Pointer: house2
Visual Salience: 0.3287

Basic Partitions 1 … *n*

VRD*n-1*
Domain Name: X
Profiled Element List: Null
Partitions: TYPE & Basic

…

Discourse Context

LRD1

Domain Name: house

Profiled Element List:
house1

Type Partition.
Diff Criterion: house

Partition Element List:
Element 1:
Object Pointer: house3
Visual Salience: 0.5349
Element 2:
Object Pointer: house2
Visual Salience: 0.3287

Basic Partitions 1 … *n*

Visual Salience Module

World Model

A Local/Episodic Architecture

- The structure of the LIVE perceptual memory and discourse context components is somewhat similar to the episodic Buffer in Baddeley's model:
    1. limited capacity,
    2. chronologically ordered,
    3. integrating visual perceptual information with semantic information

A Local/Episodic Architecture

- The similarity in the encodings in the perceptual memory and discourse context model facilitates reference resolution, which entails copying, restructuring, and inserting of a reference domains.

A Local/Episodic Architecture

- The approach to resolving a reference taken by the LIVE system can be understood as:
  1. searching memory for a suitable episodic memory,
  2. using this episode as local context within which the reference is resolving,
  3. updating the episode to mark the fact that the reference has occurred,
  4. updating the discourse context by storing the new episode in it.

A Local/Episodic Architecture

- LIVE can process exophoric references to entities that are no longer on-screen.
- However, using a reference domain representation of a frame/episode as defining the (local) context for a reference makes it extremely difficult to handle references that refer to two or more entities that never appeared in the same frame.

A Local/Episodic Architecture

- Handling references to entities perceived in different episodes requires the system to be able to integrate multiple reference domains, and this is non-trivial; e.g., it is not clear how salience scores from different frames, and hence different times, should be updated during this merger.

A Global/Monolithic Architecture

- An approach to the design of a perceptual memory, that naturally answers the question of how to integrate information from perceptions received across distinct times, is to use an evolving global structure where all referents are stored in a single data structure that is continuously updated to reflect the current state.

A Global/Monolithic Architecture

- Kelleher (2006) is similar to Kelleher et al. (2005) in that it uses the same visual salience algorithm to analysis the visual frames the user sees as they navigate through the environment.
- However, the data structure used to store perceptual memories and discourse structure is very different.
- This system maintains a single global context model throughout a user's session.

A Global/Monolithic Architecture

- Once an entity has been rendered on screen a representation of that entity is introduced in this global context model.
- There is only ever a single representation of an entity in the global context model.

A Global/Monolithic Architecture

- This representation of an entity stores:
  1. the physical information of the entity (e.g., *type*, *colour*, *size*, and so on)
  2. the current visual salience salience score
  3. the current linguistic salience score

A Global/Monolithic Architecture

- The visual salience score is updated after each frame is processed.
- The visual salience of an entity that is not in the current frame is halved when the frame is processed.
- As a result the visual salience of an entity drops off once it goes out of (visual) focus (i.e., off-screen), and continues to reduce the longer out of focus it remains.

A Global/Monolithic Architecture

- The linguistic salience scoring is based on the assumption that entities that have been mentioned recently are more salient than entities that have not.
- The particular function used to calculate and update the linguistic salience scores is in the spirit of Centering Theory Grosz et al. (1995) and is similar to the model proposed by Krahmer and Theune (2002).
- Other linguistic salience models could easily be switched in, for example (Kennedy and Boguraev, 1996) which was used in the Companion's project (Smith et al., 2010), would also be suitable.

A Global/Monolithic Architecture

- Let U_i be a sentence uttered in state s_i, in which reference is made to $\{d_i, \ldots, d_n\} \subseteq D$.
- Then the salience weight of objects in s_{i+1} is determined as follows:

$$sf(s_{i+1}, d) = \begin{cases} 1 & \text{if } d = subject(U_i) \\ (sf(s_i, d)/2) + .5 & \text{if } d = object(U_i) \\ (sf(s_i, d)/2) + .25 & \text{if } d = other(U_i) \\ sf(s_i, d)/2 & \text{if } d \notin \{d_i, \ldots, d_n\} \end{cases}$$

A Global/Monolithic Architecture

- The linguistic salience of an entity is updated after each utterance has been processed.
- The linguistic salience of any entity not mentioned in an utterance is halved when the utterance is processed.

A Global/Monolithic Architecture

- The structure of the global context model itself is minimal, it is simply an unordered set of these entity representations.
- The fact that the linguistic and visual salience scores are updated based on recency of being visible or mention means that the context model implicitly models recency.

A Global/Monolithic Architecture

- Reference resolution in this system is done by calculating an integrated salience score for each entity in the context model, and then selecting the entity with the highest integrated score as the referent.
- The integrated salience score of an entity is recalculated each time a referring expression is processed.

A Global/Monolithic Architecture

- The integrated salience score is calculated in three steps:
  1. a reference relative visual salience score is calculated by scaling the standard visual salience score to reflect the fit of the entity with the selection restrictions specified in the expression
  2. a reference relative linguistic salience score is calculated in a similar way to the reference relative visual salience score;
  3. the integrated salience score is then calculated using a weighted sum of the reference relative visual and linguistic salience scores, where the weighting is dependent on the surface form of the referring expression

A Global/Monolithic Architecture

- The fact that this monolithic global context model does not encode an episodic (frame based) structure means that the integration of information from different scenes is straight forward.
- As a result, this system can handle references to entities that do not appear on screen together.

A Global/Monolithic Architecture

- However, the loss of the episodic chronological order means that a system using this context model would not be able to handle exophoric references based on:
  1. chronology (such as *the first blue house we saw*),
  2. co-occurrence within a local temporal context (such as *the car that was in front of the house when the man fell*).

Summary

Summary

Things not mentioned:

- Reformulation of referring expressions and perception (Schütte et al., 2017; Schutte et al., 2015, 2014)
- Generating Referring Expressions: (Dale and Reiter, 1995; van Deemter, 2002; Kelleher and Kruijff, 2005, 2006; Deemter et al., 2012)
- Grounding image captioning using image retrieval (e.g. perception) in order to generate more diverse and meaningful captions (Lindh et al., 2018)
- See also cognitive architectures such as ACT-R (Anderson, 2009), SOAR (Laird, 2012)

Summary

- The two approaches to perceptual memory we have reviewed can be understood as exemplars at opposing ends of a spectrum of design choices:
  1. Kelleher et al. (2005) focuses on identifying a local context and resolving the reference within that context,
  2. Kelleher (2006) focuses on creating and continuously evolving a global context model.

Summary

- These approaches have complementary strengths and weaknesses.
- Consequently, it is likely that a blend of these approaches is necessary.
- This is not surprising as there are many examples in language processing where there is a need to be able to switch from a local focus to a global perspective, and back again, as the context requires.[3]

[3] Switching between local and global representations, similar to the challenge of modelling long-distance dependencies in sequential data Mahalunkar and Kelleher (2018)

References I

Alshawi, H. (1987). *Memory and Context for Language Interpretation*. Cambridge University Press.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?*, Volume 3. Oxford University Press.

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge University Press.

Atkinson, R. C. and R. M. Shiffrin (1968). Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation*, Volume 2, pp. 89–195. Elsevier.

Baddeley, A. D. (2002). Is working memory still working? *European psychologist 7*(2), 85.

Byron, D. (2003). Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University.

References II

Clark, H. H., S. E. Brennan, et al. (1991). Grounding in communication. *Perspectives on socially shared cognition 13*(1991), 127–149.

Coradeschi, S. and A. Saffiotti (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems 43*(2-3), 85–96.

Craik, F. I. and R. S. Lockhart (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior 11*(6), 671–684.

Dale, R. and E. Reiter (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science 19*(2), 233–263.

Deemter, K. v., A. Gatt, I. v. d. Sluis, and R. Power (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science 36*(5), 799–836.

Dobnik, S. (2009, September 4). *Teaching mobile robots to use spatial words*. Ph. D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, UK. 289 pages.

References III

Dobnik, S. and J. D. Kelleher (2016, July 16–18). A model for attention-driven judgements in type theory with records. In J. Hunter, M. Simons, and M. Stone (Eds.), *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, Volume 20, New Brunswick, NJ USA, pp. 25–34.

Eysenck, M. W. and M. T. Keane (2013). *Cognitive psychology: A student's handbook*. Psychology press.

Gorniak, P. and D. Roy (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research 21*, 429–470.

Grosz, B. (1977). *The Representation and Use of Focus in Dialogue Understanding*. Ph. D. thesis, Standford, University.

Grosz, B., A. Joshi, and W. Weinstein (1995). Centering: A framework for modelling local coherence of discourse. *Computational Linguistics 21*(2), 203–255.

Grosz, B. and C. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics 12*(3), 175–204.

References IV

Hajicová, E. (1993). *Issues of Sentence Structure and Discourse Patterns*, Volume 2 of *Theoretical and Computational Linguistics*. Charles University Press.

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

Hawes, N., M. Klenk, K. Lockwood, G. S. Horn, and J. D. Kelleher (2012). Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the 26th Naitional Conference on Artificial Intelligence*.

Hobbs, J. (1985). On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information.

Itti, L. and C. Koch (2001). Computational modelling of visual attention. *Nature reviews neuroscience 2*(3), 194.

Kamp, H., J. Van Genabith, and U. Reyle (2011). Discourse representation theory. In *Handbook of philosophical logic*, pp. 125–394. Springer.

References V

Kelleher, J. and J. van Genabith (2003). A false colouring real time visual saliency algorithm for reference resolution in simulated 3-d environments. In *Proceedings of the Conference on Artifical Intelligence and Cognitive Science*, pp. 95–100.

Kelleher, J. and J. van Genabith (2004). Visual salience and reference resolution in simulated 3d environments. *AI Review 21*(3-4), 253–267.

Kelleher, J. D. (2003). *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. Ph. D. thesis, Dublin City University.

Kelleher, J. D. (2006). Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review 25*(1-2), 21–35.

Kelleher, J. D. (2011). Visual salience and the other one. In *Salience. Multidisciplinary Perspectives on Its Function in Discourse. Mouton de Gruyer*, Number 227 in Trends in Linguistics. Studies and Monographs., pp. 205–228. de Gruyter.

References VI

Kelleher, J. D., F. Costello, and J. van Genabith (2005, September). Dynamically structuring, updating and interrelating representations of visual and lingusitic discourse context. *Artificial Intelligence 167*(1-2), 62–102.

Kelleher, J. D. and S. Dobnik (2019). Referring to the recently seen: reference and perceptual memory in situated dialog. *arXiv preprint arXiv:1903.09866*.

Kelleher, J. D., T. Doris, Q. Hussain, and S. ONuallain. (2000). Sonas: Multimodal, multi-user interaction with a modelled environment. In *Spatial Cognition - Foundation and Applications*, pp. 171–185. John Benjamins Publishing.

Kelleher, J. D. and G.-J. Kruijff (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 1041–1048. Association for Computational Linguistics.

References VII

Kelleher, J. D. and G.-J. M. Kruijff (2005). A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Kempson, R., W. Meyer-Viol, and D. M. Gabbay (2000). *Dynamic syntax: The flow of language understanding*. Wiley-Blackwell.

Kempson, R. M. (1988). *Mental representations: The interface between language and reality*. CUP.

Kennedy, C. and B. Boguraev (1996). Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pp. 113–118. Association for Computational Linguistics.

Kievit, L., P. Piwek, R. Beun, and H. Bunt (2001). Multimodal cooperative resolution of referential expressions in the denk system. In *Cooperative Multimodal Communication*, Lecture Notes in Artificial Intelligence 2155, pp. 197–214. Springer.

Kosslyn, S. M. (1994). *Image and Brain*. MIT Press.

References VIII

Krahmer, E. and M. Theune (2002). Efficient context-sensitive generation of referring expressions. In *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CLSI.

Kruijff, G.-J., J. D. Kelleher, G. Berginc, and A. Leonardis (2006). Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pp. 343–344. ACM.

Kruijff, G.-J. M., J. D. Kelleher, and N. Hawes (2006). Information fusion for visual reference resolution in dynamic situated dialogue. In E. Andre, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber (Eds.), *Proceedings of Perception and Interactive Technologies*, Volume 4021 of *LNCS*, pp. 117 – 128.

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT press.

Lappin, S. and H. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics 20*(4), 535–561.

Larsson, S. (2018). Grounding as a side-effect of grounding. *Topics in cognitive science 10*(2), 389–408.

References IX

Lindh, A., R. J. Ross, A. Mahalunkar, G. Salton, and J. D. Kelleher (2018). Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pp. 176–187. Springer.

Mahalunkar, A. and J. D. Kelleher (2018). Using regular languages to explore the representational capacity of recurrent neural architectures. In *International Conference on Artificial Neural Networks*, pp. 189–198. Springer.

Mann, W. and S. Thompson (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pp. 83–96. Nijhoff.

McCawley, J. (1993). *Everything That Linguists Have Always Wanted To Know About Logic*(but were ashamed to ask)* (2nd ed.). University of Chicago Press.

McKevitt, P. (Ed.) (1995). *Integration of Natural Language and Vision Processing (Vols. I-IV)*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

References X

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

Schutte, N., J. D. Kelleher, and B. Mac Namee (2015). Reformulation strategies of repeated references in the context of robot perception errors in situated dialogue. In *In Proceedings of the Workshop on Spatial Reasoning and Interaction in Real-World Robotics at the International Conference on Intelligent Robots and Systems*.

Schutte, N., J. D. Kelleher, and B. M. Namee (2014). The effect of sensor errors in situated human-computer dialogue. In *Proceedings of the 1st Technical Meeting of the European Network on Integrated Vision and Language (V\&L Net) a Workshop at the 25th International Conference on Computational Linguistics (COLING)*, pp. 1–8. Association of Computational Linguistics.

Schütte, N., B. M. Namee, and J. D. Kelleher (2017). Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics 31*(5), 243–257.

References XI

Sjöö, K. (2011). *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*. Ph. D. thesis, KTH Royal Institute of Technology.

Smith, C., N. Crook, J. Boye, D. Charlton, S. Dobnik, D. Pizzi, M. Cavazza, S. Pulman, R. S. De La Camara, and M. Turunen (2010). Interaction strategies for an affective conversational agent. In *International Conference on Intelligent Virtual Agents*, pp. 301–314. Springer.

Tellex, S. (2010). *Natural language and spatial reasoning*. Ph. D. thesis, Massachusetts Institute of Technology.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics 28*(1), 37–52.

Winograd, T. (1973). A procedural model of language understanding. In R. Schank and K. Colby (Eds.), *Computer Models of Thought and Language*, pp. 152–186. W. H. Freeman and Company.