Modeling Autobiographical Memory for Artificial Intelligence Robots

by

Andrew Kope

Independent Study Project
Department of Computer Science
University of Western Ontario
London, Ontario, CANADA
April, 2012

Project Supervisor: Michael Katchabaw, Ph.D.

Abstract

The present project describes a connectionist network model for autobiographical memory grounded in psychological theory; this model improves on previous attempts to model AI (artificial intelligence) robots' event knowledge by providing a more dynamic and non-deterministic representation of autobiographical memories. A proof-of-concept implementation of the model is presented, and two demonstrations of the model's validity are provided. Possible applications of the model in AI robot design are discussed, as are the model's shortcomings and future directions.

Keywords: AI robot, autobiographical memory, connectionist network, event knowledge

Modeling Autobiographical Memory for Artificial Intelligence Robots

In computer games, AI (artificial intelligence) robot (or bot) event knowledge (functionally equivalent in this context to autobiographical memory) has traditionally been represented using one of a few, arguably naive, techniques. One technique, for example, is to use a game-time related schedule, where the AI robot's behavior and dialog advances along a pre-determined path, with progress along that path determined by checkpoints in game-clock time. Another technique is stimulus driven, where in-game events activate event knowledge objects in a predetermined master event list for a given AI bot (King, 2002).

These techniques, although conceptually straightforward, each provide a deterministic and oversimplified representation of autobiographical memory as compared to the human capability to represent event knowledge. As such, to provide more realistic AI bot performance, a new and less deterministic model of event knowledge for AI robots is required. To design such a model, I approached the problem from a psychological perspective by creating a model of autobiographical memory grounded in psychological research which can then be applied back to AI robots.

Overview of Long-Term Memory

Non-Declarative Memory

Non-declarative memory is that which cannot be verbally communicated; it is typically divided into two sub-categories, namely procedural memory and motor memory. Procedural memory covers things such as driving a car, or making toast. Motor memory includes things such as swinging a baseball bat, or riding a bicycle. Non-declarative memory, not relating to event knowledge, is therefore outside the scope of the present project.

Declarative Memory

Declarative memory, conversely, consists of the verbally communicable aspects of memory. Tulving (1972, 1987) distinguished between two types of declarative memory: semantic and episodic. Semantic memory includes knowledge of facts about the world, for example that pennies are round. Under investigation in my project is episodic memory, which pertains to knowledge of one's experienced life events.

Survey of Episodic Memory Theories

McClelland, McNaughton and O'Reilley (1995) suggested that memories are first stored as changes in the hippocampal system. They argued that these changes in the hippocampus support reinstatement of recent memories in the neocortex, that these neocortical reinstatements elicit changes in the neocortex, and that the accumulation of these changes is what stores a memory. The authors concluded with the suggestion that the hippocampal system supports rapid learning of new items (because it does not require changes to neocortical structure), while the neocortex learns slowly, thereby encoding the structure of learned experiences. Meeter and Murre (2004) later corroborated this conclusion by providing a neuropsychologically grounded account for the neocortical consonsolidation of hippocampally activated memories as described by McClelland, McNaughton and O'Reilley.

Burt et al. (1995) argued that there is evidence to support the hierarchical organization of autobiographical memory. In their view, at the top level of the hierarchy are representations of lifetime periods (e.g. time at university), at the second level of the hierarchy are representations of extended events within a lifetime period (e.g. a camping holiday), and at the lowest level are representations of specific events, or parts thereof (e.g. a specific picture). To conclude, the authors posited that the important factor for memory retrieval within the hierarchy they described is which cues are provided.

Conway and Pleydell-Pearce (2000), in a similar vein to Burt et al. (2005), argued that memories are transitory mental constructions within a self-memory system. The authors described the self-memory system as containing both an autobiographical knowledge base, and the current goals of the working self. In their model, to recall a memory a control system activates a specific part of the autobiographical knowledge base (this knowledge base is stored as a series of connected nodes in a hierarchical structure, organized by lifetime periods and themes).

Williams, Conway and Cohen (2008), departing somewhat from previous theories on autobiographical memory, argued for the script theory of memory. In their view, memories are stored as scripts of behaviors and outcomes, structured within the memory system. These scripts include roles, goals, and subscripts, relevant/irrelevant actions, and are constantly reorganized following new experiences. Scripts within the memory system are assembled as required and new experiences are built from existing memories, while common elements among scripts are represented on a higher level. Agreeing with previous research (e.g. Burt et al., 2005), in their view memory retrieval is elicited by cues.

Özteckin, Davachi and McElree (2010) performed an fMRI study using a recognition task to compare hippocampal activation during working memory versus long term memory retrieval tasks. Based on their results, they argued (contrary to traditional views) that there is actually a dichotomy between focal attention and memory representations, rather than the classical distinction between long term memory and working memory representations.

Key Features

In reviewing the aforementioned papers, several key features have emerged which my model combines to create a reasonable model for human autobiographical memory grounded in psychological theory. These key features are:

1. Strengthening over time of memories with repetition (McClelland, McNaughton & O'Reilley, 1995; Meeter & Murre, 2004).
2. Cues provided are a dominating factor and elicit memory retrieval (Burt et al., 1995; Williams, Conway, & Cohen, 2008).
3. A memory is activated by a control system, which activates a specific part of the autobiographical knowledge base (Conway, Pleydell-Pearce, 2000).
4. The memory system is dynamic – structures are assembled as required and new experiences are built from existing memories (Williams, Conway, & Cohen, 2008).
5. One store handles all types of memory (Özteckin, Davachi, & McElree, 2010).

The Present Model

Description

To build a model of autobiographical memory possessing all of the key features as described above, I used a connectionist network to represent memories. A memory pattern, in the context of my model, is a collection of keywords (features), and a textual description of the memory. For example:

```
Going on vacation with Dad to Scotland
dad
scotland
vacation

Fishing at the lake by the cottage on vacation
vacation
fish
lake
cottage
```

A memory pattern is stored within the network as a collection of interconnected nodes, with each node representing a key feature of the memory (e.g. vacation). Additionally, as shown in Figure 1, each node of a memory pattern is connected to every other node in that pattern (key feature five).

Building the Network

The network is built by a control system, which reads in one memory pattern at a time (key feature four). First, they keywords are parsed; if one of the pattern's keywords is not already represented as a node in the network, a node for that keyword is created (duplicates are not allowed). Second, connections between nodes are built. Each node is connected unidirectionally to every other node in its pattern, so any two connected nodes are connected bidirectionally.

As the control system parses a pattern, if a connection between two nodes already exists its strength is increased (key feature one); this allows a pair of nodes to become more strongly connected if many memories relate them. This increase in connection strength with multiple presentations is consistent with current views on cortical rewiring and information storage in the memory system (e.g. Frankland & Bontempi, 2005; Chklovskii, Mel, & Svoboda, 2010).

Shown below is what the model's network would look like after adding the two example patterns above. For simplicity, this example uses no duplicate connections, and so the default weight of 0.8 for each connection is omitted from the diagram.

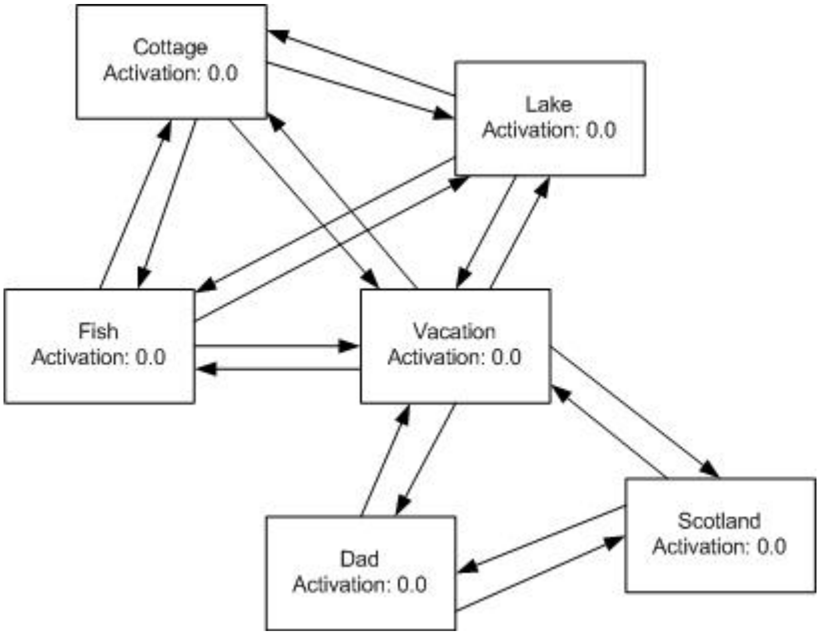

Figure 1. The network after it has been built.

Activating the Network

When provided a cue, the network controller activates the relevant nodes serially (key feature two and three). When a given node is activated, its activation increases as a weighted average of the incoming activation and its current activation. The activated node then spreads a fraction of its new activation to all the nodes it is connected to. As such, by activating one node, all of its "neighbors" are activated, and so on until the activation has spread to all connected nodes. Shown in Figure 2 below is the activation of each node after the cue *dad* is provided.

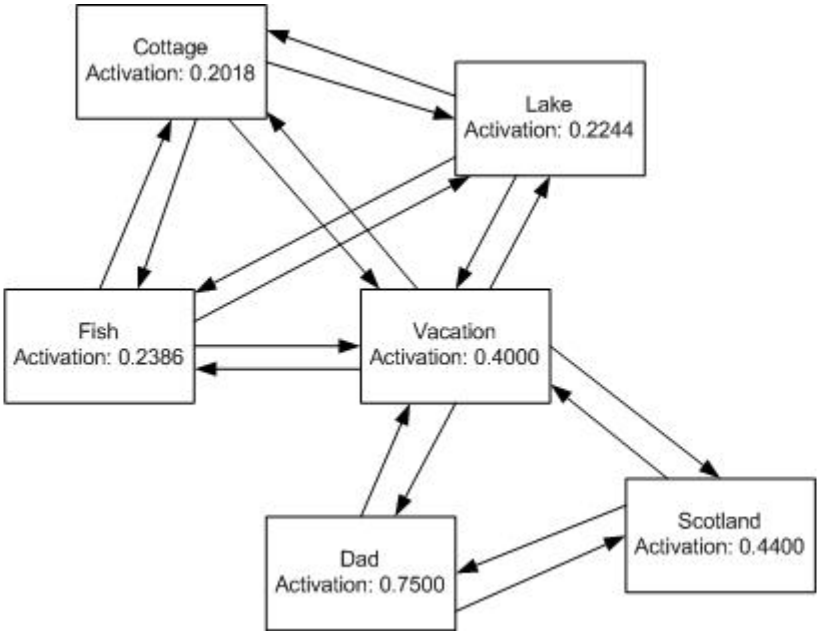

Figure 2. The network after *dad* is cued.

Passage of Time

To account for the passage of time within the network, each node's activation is decreased according to a bounded exponential function on a set time interval (this decrease is called a 'tick'). The function used is based on the forgetting curves described in Sikström (2000).

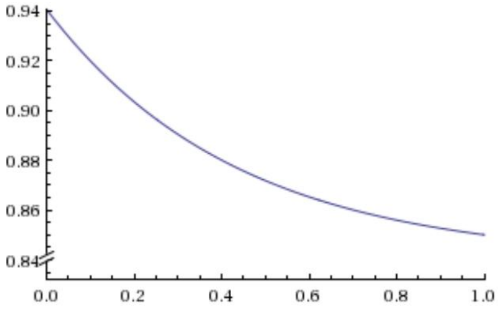

Figure 3. Rate at which activation is decreased.
x axis: Current Activation of Node
y axis: Activation Multiplier

Shown in Figure 4 below is the example network after one tick. Nodes which were more active (e.g. *dad* at 0.7500) decreased in activation more than nodes which were less active (e.g. *cottage* at 0.2018); that is by 0.1066 and 0.0200 respectively.

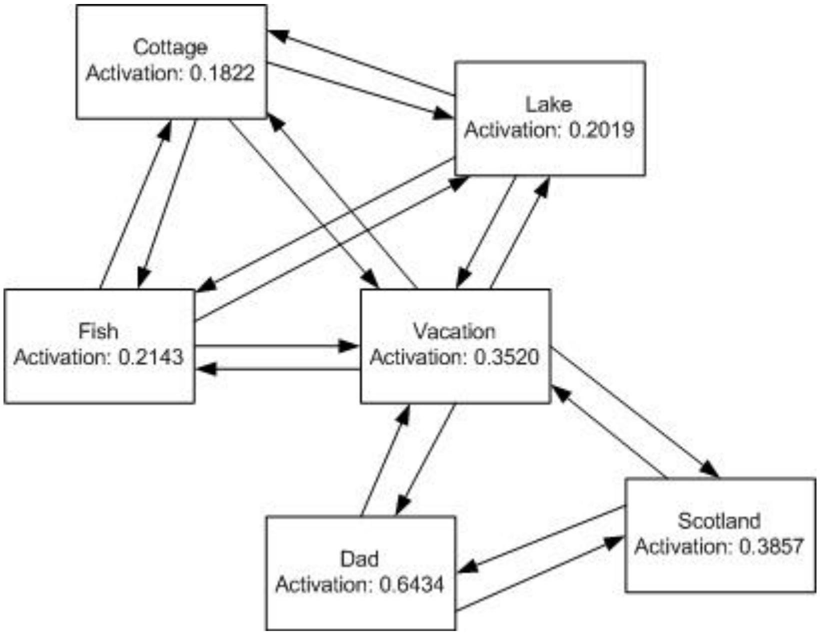

Figure 4. The network after one time tick.

Memory Retrieval

As shown above, a given cue elicits a specific pattern of activation in the nodes of the network. To quantitatively assess which memory pattern this activation represents, the n most active nodes in the network are compared to the database of memory patterns to determine which pattern has the most keywords (by percentage) in common with the n most active nodes. The pattern with the most keywords in common is deemed to be the recalled memory.

Continuing the example above, with n = 4 (Figure 5 shows the four most active nodes are highlighted in red) we have a 67% match with the memory *Fishing at the lake by the cottage on vacation*, and a 50% match with the memory *Going on vacation with Dad to Scotland*, so the network would report that it recalls the latter memory.

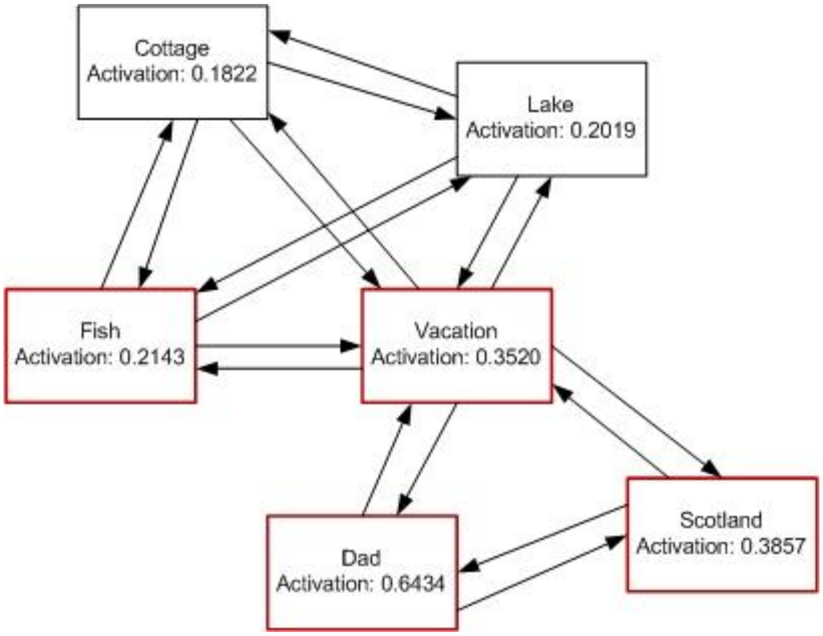

Figure 5. The n most active nodes with n = 4.

Prototype

I implemented my model in Java using the Eclipse development environment. All random values used by the model are pseudorandom values provided by the standard Java library Math.random() function. Node and Connection objects are stored in hash tables for ease of access and storage, while memories are stored in a vector. The activation value for each node is stored using a double precision floating point value, and all calculations of activation change are carried out in double precision. Memories are stored as text documents and are parsed by a control system to build the network.

Model Assessment

To assess the model, I tested its ability to demonstrate two properties of human autobiographical memory recall; that context affects which memory is recalled given a specific cue, and second that given an initial cue, the model is capable of performing a ‘free recall’ task producing a series of related memories.

56 memories were parsed for the purposes of the tests, with between 3 and 5 keywords each. After building the network, there were a total of 86 unique nodes with 682 unidirectional connections.

Context Dependency Demonstration

Shown below is the effect of context on the activation resulting from the cue *restaurant*. The activation of the seven most active nodes in the network and the corresponding memory pattern activated is displayed after each cue is presented.

```
-------------------------------------------------------------------------------------
Current pattern number: 0
*****Current Pattern:
friend
house
restaurant

*****After input keyword "friend" :
Memory:
friend  0.75
school  0.5353085695152514
football  0.5176928988817009
game  0.5168490236347945
gym  0.5122515823440711
lab  0.5120126558855028
party  0.5099118502392075
Playing a game of squash at the school gym with a friend

*****After input keyword "house" :
Memory:
house  0.8569341393142164
friend  0.6644873362447593
school  0.6188513119018506
party  0.609227024932279
cake  0.6012362429580752
laptop  0.5955691641091232
girlfriend  0.5947115514639376
Having cake at girlfriend's birthday party at her house with friends

*****After input keyword "restaurant" :
Memory:
restaurant  0.8552614249707078
house  0.7318180878052285
dinner  0.6593790103051937
girlfriend  0.6371149829129985
granddad  0.618033091650236
steak  0.6147112960280383
dad  0.6113219545202668
Going out for dinner with girlfriend to a restaurant, had steak and wine

-------------------------------------------------------------------------------------
Current pattern number: 1
*****Current Pattern:
```

```
grass
icecream
restaurant

*****After input keyword "grass" :
Memory:
grass  0.75
sun  0.4759654431680001
accident  0.47234739669594406
weekend  0.4636900902400001
icecream  0.4577618377733994
frisbee  0.4426658432000001
school  0.434685486245467
Playing frisbee in the park on the grass on a sunny weekend

*****After input keyword "icecream" :
Memory:
icecream  0.8501455555350781
grass  0.6617719027331039
accident  0.5940415253682287
school  0.5754216279410012
granddad  0.5747218640204472
restaurant  0.5740847889763457
spill  0.5737651559048645
Dropping ice cream on the grass at school by accident

*****After input keyword "restaurant" :
Memory:
restaurant  0.8744104654968425
icecream  0.7303564529492172
dinner  0.6546585596088546
granddad  0.6542213203404905
spill  0.6328895861185541
necktie  0.6324309303555855
hockey  0.6320918880427464
Spilled ice cream on my necktie at a restaurant with granddad

-------------------------------------------------------------------------------------
Current pattern number: 2
*****Current Pattern:
dad
hockey
restaurant

*****After input keyword "dad" :
Memory:
dad  0.75
childhood  0.5015992107248093
cookies  0.5003462884756418
distraction  0.49960962937288766
brother  0.4947522750344713
smoke  0.49186587634759404
hospital  0.49123536836780624
Dad burning cookies in the kitchen

*****After input keyword "hockey" :
Memory:
hockey  0.8519310897158118
dad  0.6624861164053975
childhood  0.5983126214316805
hospital  0.597230364682277
```

```
beer  0.5824294134162364
game  0.5823065641359138
restaurant  0.580420437508208
At the hospital for a hockey wrist injury as a child

*****After input keyword "restaurant" :
Memory:
restaurant  0.8757270581497201
hockey  0.7276836558969305
dinner  0.6626522112548783
icecream  0.6494385951840322
dad  0.6433308355357668
granddad  0.6344709144026662
steak  0.6337220879707024
Going out to a restaurant for ice cream with dad after hockey
```

In the first input pattern, *friend* and *house* are the context in which *restaurant* is cued; as such the resultant memory is *Going out for dinner with girlfriend to a restaurant, had steak and wine*. In the second pattern, *grass* and *ice cream* are the context in which *restaurant* is cued; in this case the resultant memory is *Spilled ice cream on my necktie at a restaurant with granddad*. In the third input pattern, *hockey* and *dad* are the context in which *restaurant* is cued; again a different memory is recalled, this time *Going out for ice cream with dad after hockey*.

The results of this demonstration make it clear that the network activation produced by a given cue is indeed dependent upon the context in which the cue is presented.

Free Recall Demonstration

Shown below are the series of unique memories that are activated when the network performs a “free recall” task twice for the keyword *park*. The activation of the seven most active nodes in the network is also displayed after initial cue presentation, and after the free recall task.

I implemented the free recall task by giving the network a starting cue, then having the control system repeatedly randomly select one of the third to sixth most active nodes in the network and then randomly activate that node either once or twice. Each time a new node is activated, if that activation caused a new memory to be recalled, that memory is displayed. New nodes are activated until five unique memories are recalled (the number of activations required is displayed as “Duration of reflection”).

```
*****After input "park" :
Memory:
park  0.7815784428720013
game  0.5689221006315489
sidewalk  0.5655802181255364
leash  0.5654088415159119
friend  0.5654052599624169
grass  0.5654044532787267
dog  0.5653999919945515
Taking the dog for a walk on a leash on the sidewalk in the park

*****Free Recall:
Taking the dog for a walk on a leash on the sidewalk in the park
Playing a game of football with friend in the park
Working on a project with a friend in the school lab
Having beer with a friend at a bar with music and dancing
Watching a hockey game on tv with a friend, drinking beer
Duration of reflection:7
Memory:
game  0.9042592428344713
beer  0.8041248188773422
friend  0.7449972207003259
school  0.7256447680086708
football  0.7178545704967187
hockey  0.6744427420731874
television  0.6742453808254604

*****After input "park" :
Memory:
park  0.7815339559648993
game  0.568847642924267
friend  0.5653256598917844
sidewalk  0.5652926348361098
football  0.565286494552188
grass  0.5652219494665038
leash  0.5652172682078693
Playing a game of football with friend in the park

*****Free Recall:
Taking the dog for a walk on a leash on the sidewalk in the park
Stepping in spilled water on the sidewalk with shoes
Buying squash shoes at the store
Trying on pants at the store with Mom as a child
Childhood memory of Mom baking cookies in the kitchen
Duration of reflection:9
Memory:
childhood  0.9396545126239481
mom  0.7594039735262905
cookies  0.7206362123622896
hospital  0.7206306753200087
school  0.719757553987176
pants  0.7137510633273692
store  0.6989573688555899
```

As the output above demonstrates, my model is capable of producing a series of related memories given an initial memory cue. The randomness with which a new cue is selected by the control system following each cycle allows the model to produce different, though equally plausible, sequences of memories from the same cue.

AI Robot Design Applications

The connectionist network model of autobiographical memory I have presented has several applications to AI robot design. Firstly, my model provides a framework for allowing AI robots to acquire new memories over time, and integrate those new memories into their existing memory store much more flexibly than other models of AI robot event knowledge. This framework could be implemented by creating an in-game monitor routine responsible for reading the surroundings of the robot and then creating new memory patterns. The monitor could then pass the new memory patterns to network's control system which would add them to its network. This monitor routine could also encode the duration of each event in the description it produces; this would solve a shortcoming of my model, namely that it does not at present encode the relative duration of different memories.

My model could also have implications for human-AI interaction in games. For example, the model could be used to facilitate a gameplay dynamic whereby recent events and conversational context are relevant considerations for the player when influencing or predicting AI robot behavior. Importantly, this gameplay dynamic is not tractable with more naïve representations of AI bot event knowledge, which do not allow AI bots to respond dynamically and unpredictably based on recent events and conversational context.

Similarly, my model could be applied to chat bot design. A chat bot built based on this model could recall and discuss different memories depending on the conversational context. This would provide a much more complex, less predictable, and likely more realistic flow of ideas in the chat bot’s responses than a naïve user input keyword-matching algorithm.

Shortcomings and Future Directions

A major practical shortcoming of the model at present is that there is no way to distinguish between different memories which are represented with the same set of keywords. Although it is already somewhat unlikely that two different memory patterns would have the exact same keywords, this probability can be further reduced by increasing the number of keywords associated with each memory, which would render negligible the odds of two memory patterns having the exact same keywords.

Another shortcoming of the network from a theoretical perspective is that its connections do not age or decay. That is, there is no process by which the network eventually forgets old memories. Conversely however, the static nature of memory representations avoids a problem that more dynamic connectionist models of memory have; namely that well-learned information is forgotten as new information is learned (Ratcliff, 1990).

Finally, Moreton and Ward (2010) for example discuss several findings illustrating the differences between recalling recent events and distant events, however the present model has no way to capture or adjust the time scale of a memory from recent to distant once it is encoded. One possible solution for accommodating differences between the recall of recent and distant events would be to introduce a hierarchical structure to the network to represent recency, with more recent memories at the top of the hierarchy.

Conclusion

The present model captures several characteristics of human autobiographical memory described in psychology literature, namely: nodal connections strengthen with repetition, cues are the dominant factor modulating memory retrieval, memory structure construction is dynamic, and all types of memory are contained within a single store. To demonstrate its validity, the present model displayed the ability to simulate two high-level memory tasks: contextually dependent recall and non-deterministic free recall. In conclusion, the model's ability to integrate new memories into its existing network easily and its ability to recall different memories depending on context have interesting implications not only for AI robot design, but possibly also for current psychological theory on the structure and modeling of human autobiographical memory.

References

Burt, C. D., Mitchell, D. A., Raggatt, P. T., Jones, C. A., & Cowan, T. M. (1995). A snapshot of autobiographical memory retrieval characteristics. *Applied Cognitive Psychology*, *9*, 61-74.

Chklovskii, D. B., Mel, B. W., & Svoboda, K. (2004). Cortical rewiring and information storage. *Nature*, *431*, 782-788.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107*(2), 261-268.

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews*, *6*, 119-130.

King, K. (2002). A dynamic reputation system based on event knowledge. In S. Rabin (Ed.), *AI Game Programming Wisdom* (pp. 426-436). Hingham, Massachusetts: Charles River Media Inc.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419-457.

Meeter, M., & Murre, J. M. (2004). Consolidation of long-term memory: Evidence and alternatives. *Psychological Bulletin*, *130*(6), 843-857.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*(4), 510-515.

Öztekin, I., Davachi, L., & McElree, B. (2010). Are representations in working memory distinct from representation in long-term memory?: Neural evidence in support of a single store. *Psychological Science, 28*(8), 1123-1133. doi: 10.1177/0956797610376651

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*(2), 285-308.

Sikström, S. (2002). Forgetting curves: Implications for connectionist models. *Cognitive Psychology*, *45*, 95-152.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381-402). Retrieved from http://web.media.mit.edu/~jorkin/generals/papers/Tulving_memory.pdf

Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology, 6*, 67-80.

Williams, H.L., Conway, M.A. & Cohen, G. (2008). Autobiographical Memory. In G. Cohen and M.A. Conway (Eds.) *Memory in the Real World (*pp. 21-81). New York: Psychology Press.