Artificial Working Memory: A Psychological Approach

Peter Dreisiger
Maritime Operations Division, Defence Science and Technology Organisation
Department of Computer Science & Software Engineering, The University of Western Australia
email: prd@csse.uwa.edu.au

ABSTRACT

In this paper, we are going to look at how cognitive models of human memory and analogy-making can be used by artificial agents to represent context-specific knowledge, and to apply knowledge from more distant, or analogous, contexts. In particular, we will focus on an activation-based model of working memory, and a hybrid model of associative memory that is able to discover analogues and the correspondences between them.

We will also look at a model of context-mediated behaviour, and we will close by identifying how these three models can be brought together and developed into an artificial working memory system.

Categories and Subject Descriptors:

I.2.6 and I.2.4 [**Artificial Intelligence**]: Computing Methodologies — Learning; Knowledge Representation Formalisms and Methods

Keywords:

Working Memory, Associative Memory, Analogies, Context, Context-specific knowledge, Context-mediated behaviour

1. INTRODUCTION

In the field of artificial intelligence (AI), the term 'intelligent agent' usually refers to a software entity that can observe its environment and act in a goal directed manner; an agent is said to be 'rational' if it chooses actions that it believes will maximise its measure of performance, given its model of the world and its background knowledge.

According to this definition, rational behaviour ought to be quite achievable; practically, however, there are still real differences between the way *we* act, and the way in which artificial agents behave. Even in complex, real-world situations, we humans are able to infer a lot about our environment, and thus build up a detailed world model — a task that helps us to deal with uncertain, incomplete, and even erroneous information, but a task that tends to require context-specific knowledge.

In this paper, we are going to look at how cognitive models of human memory and analogy-making can be used by artificial agents to represent this type of knowledge, and to represent the contexts themselves. While this *approach* is somewhat novel, the use of contexts to capture implicit assumptions, to constrain problem-solving, and to qualify knowledge is not — previous research has looked at how we can model and represent context using logic, production rules, and local, or context-sensitive, knowledge bases (for an overview of these approaches, see [5, 20, 28]).

Indeed, the importance of context to first-order logic was discussed by John McCarthy as far back as 1979 [21]. In this paper, entitled 'Generality in Artificial Intelligence', McCarthy also touched on two other approaches to generality: the production systems of Newell and Simon focused on generalising the goal-seeking and problem-solving mechanisms, while their 'General Problem Solver' attempted to generalise the *problems* themselves, and obtained solutions by finding appropriate sets of transformations. This problem solver prefigured, at least to some extent, the structure mapping models of analogy that were developed by cognitive scientists in the 1980s [13, 15].

But why look at *these* three areas, and how do they fit together? Firstly, cognitive models of working memory give us two things: (1) a theoretical basis that is supported by empirical evidence, and (2) information about some of the idiosyncrasies of human memory that could also benefit artificial agents. Cognitive models of analogy and analogical recall tell us how to retrieve more distant memories, and they give us insights into one of *our* most important mental tools; in some cases, they even give us a general, *computational* model of associative memory. Finally, the areas of context representation and modeling give us (1) examples of how agents can benefit from a semi-autonomous context-dependence knowledge base, and (2) insight into some of the issues we might face when trying to merge contradictory knowledge.

In this paper, then, we are going to review some of the work that has come out of the field of cognitive science; in Section 2 we will look at two models of human working memory, while Section 3 will focus on models of analogy. Finally, in Section 4, we will review some models of context and context-mediated behaviour.

2. WORKING MEMORY

In cognitive psychology, the term 'working memory' refers to the structures and mechanisms that maintain task-relevant

16th School of Computer Science & Software Engineering Research Conference, Yanchep, Western Australia, 12–13 June 2008.

information in a highly accessible form for the duration of a cognitive task. In this sense, working memory is similar to its theoretical predecessor, short-term memory (STM). There are, however, some important differences, particularly in their relationship to long-term memory (LTM) and the way in which their contents are used.

Early models of the human information processing system clearly differentiated between short- and long-term memory. In Broadbent's original model ([6], as summarised in [8]), information is conveyed in a fixed order from one storage unit to the next. Percepts are first held in an unanalysed form, in a sensory store of unlimited capacity. Some of this sensory information is selected for further analysis, and the processed information is held in a limited-capacity short-term store. Finally, a subset of this information is stored in a permanent, or long-term store, which is implemented as a semantic network.

This model was not without it detractors, and by 1984, Broadbent conceded that it was too linear, and depended too heavily upon feedback loops. To address these issues, he proposed an alternative model in which the abstract working memory, sensory, long-term associative and motor output stores were arranged around a central processing system. While his now ubiquitous *Maltese Cross* model ([7], as summarised in [4, Chapter 2]) does allow information to flow freely from one store to another, it relies heavily on the processing system to translate and regulate the flow of information.

In both of these models, storage in STM is temporary; its contents are highly accessible, but when attention is diverted to another demanding task, the information stored in STM becomes unavailable in a matter of seconds. Also, the storage capacity of STM is limited to around seven (plus or minus two) items. The capacity of LTM, on the other hand, is assumed to be vast and much more durable than that of STM. In these models, its primary bottleneck is its accessibility, with retrievals from LTM taking about 1 s, and the storage of a new memory taking between 5 and 10 s.

In contrast to the traditional storage-oriented notion of short-term memory, working memory is a more active and processing-oriented construct. It has been described as the 'workspace' or 'blackboard' of the mind, and it is where the active processing and temporary storage of task-relevant information take place [23, 9]. Such a view of working memory requires a more sophisticated account of the control mechanisms that goes beyond simple memorisation strategies.

To introduce the concept of working memory, let us look at two models that typify the two main families of cognitive theories — the multiple component models, and the activation-based models. (For a good introduction to other members of these families, see [23].)

2.1 Multiple-Component models of Working Memory

The first group of models defines working memory as a plural construct that is separate from long-term memory. Multiple-component models are reminiscent of Broadbent's Maltese Cross in that they maintain information in multiple stores and are regulated by one or more executives; however, many of the models are also refinements as they include modal, or sense-specific memory and maintenance procedures.

The canonical multiple-component model was introduced by Baddeley and Hitch [2] in 1974, and is arguably the best known model of working memory. In its original form, it consisted of three components: the *central executive*, and two auxiliary, or 'slave' systems called the *phonological loop* and the *visuospatial sketchpad*. The central executive is responsible for coordinating the slaves and directing the focus of attention; the slave systems provide temporary storage for verbal and visual information, respectively. This partitioning of working memory was based upon the selective interference effects found in dual-task tests, and the impairments associated with specific types of brain damage [3].

According to this model, storage in these slave systems is volatile, and without active rehearsal their contents will become inaccessible within seconds. Information in the phonological loop is maintained using processes similar to silent rehearsal; they were less clear on how visual and spatial information is maintained. These processes were made more explicit by Baddeley and Logie [3] who further divided each of the stores into a passive memory, and an active rehearsal component that is under the control of the executive.

The central executive, too, has undergone some changes. Originally, it included some short-term storage of its own that could supplement the slave systems or store memory traces associated with the other senses; more recent accounts of the multiple-component model have dropped this assumption in favour of an episodic buffer and access to long-term memory [3]. □

These models are consistent with our ability to maintain verbal and visual information, and they account for the interference effects seen in studies and in everyday life. However, I do not believe that they should form the basis of an *artificial* memory system.

Firstly, their focus is on how we maintain *perceptual* information and, to a lesser extent, how we prepare for specific types of actions; they do not really account for how we maintain, and manipulate, declarative knowledge. Their central executive poses another problem; as Baddeley and Logie said of their own theory [3, p. 39], one problem with a "control structure like the central [executive] is that such a model simply postulates a homunculus, a little person who makes all the awkward decisions in some unspecified way and, hence, that it adds nothing in explanatory value".

2.2 Working Memory as an Activated Subset of Long-term Memory

The other group of models treats working memory as an activated subset of long-term memory that is governed by a single, or mode-independent, mechanism and includes the theories of Cowan [9] and Ericsson and Delaney [10]. While this view of working memory dates back, in part, to Norman [24], it was not until Cowan's review of 1988 [8] that its implications upon the flow of information, the role of the central executive, and the mechanisms of selective attention were considered in detail.

A full review of Cowan's arguments and evidence is beyond the scope of this paper, but we will look at the key points of his *Derived Components of Processing* model of the human information processing system and its successor, the *Embedded-Processes* model of working memory. The essence of these models is that (1) working memory is hierarchical, (2) we habituate to, rather than filter out unwanted stimuli, (3) attention and awareness are directed by both voluntary and involuntary processes, and (4) awareness affects the way in which memories are encoded and retrieved.

Cowan's models of working memory consist of long-term memory, the activated subset of memory (which equates roughly to short-term memory), and the focus of attention. Traces in long-term memory can be activated by stimuli, priming or voluntary processes, but only representations with a sufficiently high level of activation may enter the focus of attention. While Cowan doesn't identify any limits to the total *amount* of activation, he notes that, without maintenance, activation tends to fade within 10 to 20 s; in contrast, he suggests that the focus of attention is capacity limited and can only hold around four active items at a time.

On the second point, Cowan argues that it is habituation, and not filtering, which directs our focus of attention. Instead of *blocking* an unattended stimulus, the processing system develops a model of its physical characteristics. This model leads to habituation which, in turn, suppresses further processing of the unwanted stimulus, thus allowing representations of habituated stimuli to be activated without ever entering awareness. The habituation hypothesis is also consistent with observations that are much harder to explain using traditional filter-based models of attention (the most notable of these is that physical changes in an unattended stimulus are easy to detect, but semantic changes are much harder to recognise).

Cowan's third point is that attention is directed conjointly by voluntary and involuntary processes. According to his theories, a memory may be activated through effortful processes, or by exposure to a stimulus. On this, he wrote "If concepts 'rise to active attention' by virtue of the total activation resulting from automatic and attentive sources together, then it might also be possible for a concept to reach awareness because its automatic activation alone surpasses a certain level" [8, p. 171]. While this may seem fairly intuitive, involuntary processes, and the role they play in directing our attention, seem to have been largely ignored by many multiple-component models.

Voluntary attention is directed by a central executive. As in Baddeley and Logie's model [3], the executive directs the focus of attention and activates representations within long-term memory; however, Cowan's central executive is also responsible for maintaining information in short-term memory, and for increasing the efficiency of memory coding and access.

On the involuntary processes, Cowan identifies three situations in which a stimulus in an unattended channel may draw resources away from the prior, voluntary focus of attention. They are: (1) when there is a change in the physical characteristics of the unattended stimulus, (2) when the stimulus is of personal significance to the subject, and (3) when an unattended channel contains information that has been primed by recent context.

His final point is that attention affects the extent to which memory traces are encoded, and the way in which they are retrieved; "In perception it increases the number of features encoded, and in memory it allows new episodic representations to be available for explicit recall" [9, p. 65]. He also argues that memories encoded with effort and awareness are easier to recall because they allow us to incorporate "contextual constraints that would not automatically be taken into account" [8, p. 176]. □

While the multiple-component models focus on sense-specific information, Cowan provides a general, domain independent model of memory and the human attentional system — that is, his theories focus on the structure of working memory, the mechanisms of selective attention, and how memories are activated.

More importantly for us, Cowan's theories are well suited to computer modeling. His models are consistent with the ACT* and ACT-R models of Anderson and Lebiere [19], and the distinction between voluntary and involuntary processes complements ACT-R's concepts of base-level (context independent) and source (contextual) activation. There are also parallels between his central executive and the pragmatic unit of Holyoak and Thagard's Analogical Constraint Mapping Engine [15].

However, it is the relationship between Cowan's theory and the Associative Memory-Based Reasoning (AMBR) model of Kokinov and his colleagues [17, 18] that is particularly interesting. Like Cowan's theory, AMBR's model assumes that (1) working memory is hierarchical, (2) focus can be directed by internal and perceptual processes, and (3) the retrieval process is affected by the reasoner's awareness, or internal context[1]. (Section 3.2 provides a more detailed review of AMBR and its associated studies.)

2.3 Artificial Working Memory

Thus far, we have focused on theories of human memory — many of which have been implemented as computer models. These models have been used to further our understanding of the human mind; they encourage rigour, they allow us to validate hypotheses by generating predictions that can be compared with observed data, and they can be used to systematically compare different theories.

In the area of memory research, computer models have been used to assess models of the phonological loop, and to investigate predictions of serial recall performance and intrusion errors. They have also been used to investigate temporal

[1]Despite these similarities, there are also some key differences: (1) AMBR is a model of analogical reasoning, and not just of working memory, (2) AMBR is a hybrid model in which each element of LTM has both an activation and a symbolic processor, and (3) AMBR's architecture is fundamentally decentralised — concepts and episodes are represented by *coalitions* of elements, and the task of the executive is itself distributed across the active elements of working memory.

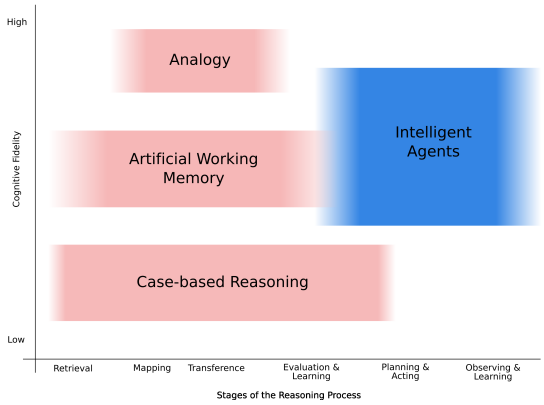

Figure 1: The relationship between analogy, case-based reasoning, artificial working memory and intelligent agents.

and contextual dependencies in LTM, and to model serial recall in working memory. The AMBR family of models have even been used to study the effects of context and priming on recall, problem-solving, re-representation and episode blending [18].

We have seen that working memory plays a crucial role in cognition, and that computer models of working memory can help us to further our understanding of the human mind. But why would we want to build an *artificial* working memory system? There are several reasons. Perhaps the most compelling reasons come from its ability to separate retrieval from reasoning, and its ability to acquire new concepts.

When designing an artificially intelligent agent, we need to consider many things — how will we represent the environment? what kind of background knowledge should we give it, and how will it access this information? how will it cope with uncertainty and incomplete data? how will it choose its actions? and how will we rate its performance?

By placing working memory between an agent's sensors and its decision making element, we can give it the ability to recognise existing contexts, and reason using precedents — even analogies. This, in turn, allows the designer to focus on the agent's heuristics. (Turner et al. adopted a similar approach in developing their Context-Mediated Behaviour [28, 1]; see Section 4.2 for more details.) And if our memory system is able to learn at the sub-symbolic level (i.e. it can discover, and *represent*, significant associations and structures), then we might even be able to give the agent new *symbols* to reason about.

Another reason to create an artificial working memory system is that doing so will also give us a framework within which to investigate different types of similarity, measures of uncertainty, and knowledge bases (independently, or when connected to an agent). The final reason is less concrete, but important nonetheless. By alluding to its natural counterpart, the term *artificial working memory* encourages us to consider and incorporate the research of cognitive scientists.

For an artificial memory system to be of use to an agent, though, I believe that it needs to:

1. *increase the observability of the environment* by storing recent percepts and inputs;
2. *identify relevant background knowledge and precedents*, and make that information available to the agent;
3. *make inferences about the current environment* by drawing upon similar situations, episode fragments, analogues and schema;
4. *be able to maintain multiple hypotheses* and change its assumptions and inferences as the agent's goals and environment develop;
5. *be able to discover new associations, concepts and contexts*, and generate new symbols that represent their essential features; these new concepts could then be used, by the agent or the memory system, to classify settings or episodes;
6. *be able to operate continuously.* If we ignore their subtleties for a moment, the aim of most models of analogy and case-based reasoning is to retrieve, adapt and apply the episode description that most closely parallels the current situation — that is their processing starts with a snapshot of the current situation and ends when they have produced an 'answer'. Working memory, on the other hand, doesn't take a single input, nor does it have an end product — rather, its inputs are continuous streams that represent the environment and the agent's goals;
7. *interact with the agent*; that is, *it* should be able to affect the agent's focus of attention, and the agent should be able to change the distribution of activity in working memory;

The first two of these features are quite unremarkable and are usually implemented, at least to some extent, in the agents themselves. The ability to make inferences from general knowledge and precedents, both closely related and more distant, is less common; and being able to represent, and distinguish between, different contexts is even more novel. However, what differentiates *artificial* working memory from the models of analogy and case-based reasoning is the concept of a memory system that sits between an agent's sensors and its decision making element *and* interacts with the agent's logic.

3. ANALOGICAL REASONING

Having looked at the concept of working memory, and Cowan's activation-based model in particular, we are now going to turn our attention to the area of analogies, and how they can be used to recall more distant memories and precedents.

Analogy is both the resemblance of relations, and the cognitive process of transferring knowledge from one concept, or domain, to another. Analogies enable us to learn and to reason about new concepts in terms of the familiar; we use them to draw parallels between dissimilar situations; and they allow us to solve specific problems by applying our knowledge of general principles.

In the paper which introduced the *Structure-Mapping* theory of analogy, Gentner identified several different kinds of comparisons and defined them in terms of the types of predicates they map [13]. In particular, she defined *literal similarity* as a comparison which maps attributes and relations from a source to the target, *analogy* as a comparison which maps primarily relations, and *abstraction* as an analogy which relates the target to a system of abstract or variable-like concepts.

While these definitions of similarity and analogy are widely accepted, there is less agreement on what it means to 'reason by analogy'. Most models of analogical reasoning identify the following stages [29, 13, 27, 17]:

1. a *retrieval* or *access* stage, which searches for sources that are similar, or analogous to, the current target;
2. a *mapping* stage, which finds, and applies, one or more mappings from the source analogue to the target;
3. a *transference* stage, in which these mappings are combined with knowledge about the source domain to make additional inferences about the target;
4. an *evaluation* stage, which estimates the quality of the analogy based upon the types of its mappings and how well their inferences agree with the known facts; and
5. a *learning* stage, which incorporates new information into its knowledge base.

When it comes to the details, however, there is less consensus. Some theories try to account for all of the stages [29, 17] while others focus on just one or two of them. Some *require* the source and target analogues to come from different domains [13] while others simply define 'reasoning by analogy' as the transfer of knowledge from object to another. Of those theories that provide an account of the retrieval stage, some take an agent's context and goals into account when searching for mappings [27, 17], while others consider only the source and target analogues.

In this section, we are only going to look at two models of analogy and analogical reasoning — the first is arguably the best known *traditional* model of analogy, and the other is a hybrid, multiple-constraint model. There are, of course, many others (including the early works of Patrick Winston [29], and ACME and ARCS [15, 27], which were the first models to consider the multiple constraints of structure, semantics and pragmatics), however these have been omitted for the sake of brevity.

3.1 The Structure Mapping Theory

The *Structure-Mapping* theory of analogy was proposed by Dedre Gentner in 1983 [13] and implemented in the *Structure-Mapping Engine* (SME) of 1986 [11]. While this was not the first attempt to model the process of analogy on a computer, Gentner was the first person to develop a formal theory of mapping and transference based upon systems of interconnected relationships.

In structure-mapping, knowledge is stored in a propositional network. *Concepts* are represented by nodes, *attributes* correspond to single-argument predicates, and *relations* are represented by predicates of two or more arguments. Predicates are further classified as first-order or higher-order, depending on whether their arguments are concepts or other predicates.

Her theory also defines three types of similarity. *Literal similarity* involves the mapping of attributes and relations. In an *analogy*, only relational structures are mapped, and an analogy which maps an abstract description to a target concept is called an *abstraction*.

The central ideas of her theory are: (1) that analogies are relation-preserving mappings between concepts or domains; and (2) that systems of nested relations should be favoured over isolated predicates because they describe causal chains or constraining principles. The second idea, which she calls the *systematicity principle*, determines which relations should be mapped from the source domain into the target.

Psychological studies by Gentner et al. appear to support this principle of systematicity. They found that adults do indeed focus on shared systematic relational structures when interpreting analogies and judging their soundness [12], and that systematicity can also account for *some* aspects of context sensitivity. □

Even though the structure-mapping theory may be an intuitive model of the way we reason by analogy, it has several limitations. Firstly, by itself, it does not address the retrieval stage of analogical reasoning; rather, it assumes that suitable source analogues have already been found. (A model of similarity-based retrieval, called MAC/FAC, *was* developed to support the SME, however, the combined system still only implements a sequential model of analogy — i.e. MAC/FAC is essentially a pre-processor.)

Secondly, while the SME tries to find those mappings that best preserve the relational structures of the base and target, it requires a literal transfer of relations; unlike other theories (which consider predicates to be similar if they are synonyms, hyponyms or meronyms), the original structure-mapping engine will only match relations if they are identical.

And finally, the SME emphasises systematicity and structural consistency over pragmatic and semantic constraints. This is in contrast to other models (such as ACME [15] and AMBR [17]) which use the systems' goals to identify particularly salient features, and the relationships described in their knowledge bases to support semantic *and* structural similarity.

3.2 Associative Memory-Based Reasoning

Associative Memory-Based Reasoning (AMBR) was first introduced by Boicho Kokinov in 1988, however the version we are interested in is based upon the hybrid models described in 1994 [17]. (AMBR was developed to model the spontaneous use of analogy in problem solving, it has also been used to investigate the effects of priming and context, and episode-blending and re-representation.)

While other hybrid models of analogy have been proposed over the years (see, for example, [15, 27]), AMBR is unique in several respects: (1) it is inherently dualistic — i.e. each

fundamental processing unit is a symbolic processor *and* part of a connectionist system; (2) AMBR's behaviour and characteristics emerge from the interactions of its agents — i.e. it has no central executive; (3) it finds analogies by running the retrieval and mapping processes *concurrently*; and (4) it represents episodic knowledge in a decentralised manner.

The DUAL Cognitive Architecture

To understand how AMBR works, and to see how its symbolic and connectionist components interact with each other, we first need to look at its underlying cognitive architecture. DUAL is a multi-agent architecture that is built around large networks of simple hybrid processing units, called *micro-agents*[2] True to its name, much of the architecture can be described from two different perspectives. DUAL itself can be thought of as a memory system or a parallel distributed processor; the agents can be thought of as representational or processing units; they represent both symbolic and associative knowledge; and the links between the agents have symbolic *and* connectionist significance.

If we think of DUAL as a memory system, then the total set of agents constitutes its long-term memory (LTM), while the set of active agents makes up its working memory. Each agent represents a specific piece of knowledge, and *every* agent has a symbolic and a connectionist component. If the symbolic part of an agent represents a piece of declarative knowledge, then the connectionist part represents its relevance, and thus its accessibility in the current context. If, on the other hand, the symbolic component represents procedural knowledge, then the agent's activation determines whether the operation is allowed to proceed and its rate of execution [26].

DUAL can also be thought of as a parallel distributed processor, where each agent maintains a small, local store of information, and can perform a few simple symbolic and connectionist operations. Practically, DUAL agents are implemented in LISP as frame-like structures. While they are often used to represent objects, concepts, propositions, situations or rules, they can also define daemons and procedures — this is how, for example, spreading activation and marker-passing are implemented.

In fact, these two mechanisms are built into every agent, and it is through them that the phenomenon of associative memory emerges. Where spreading activation uses associative relevance to reduce the size of the search space, marker-passing uses causal relevance and semantic similarity. In AMBR, marker-passing is used for two reasons: (1) to see if two agents share a common super-class, and (2) to find a correspondence between the elements of two sets of agents.

Links, too, serve a dual purpose. From a symbolic point of view, a link can represent an arbitrary relationship; the architecture does, however, define several standard link types including sub-class, super-class, instance and instance-of. From the connectionist perspective, each link has a weight that reflects the strength of an association between two agents, and which determines how much activation will be shared between them. These links may be permanent or temporary; excitatory or inhibitory. The permanent ones come from LTM and are excitatory, while the temporary links are created by agents to enforce specific constraints — these links may be excitatory or inhibitory.

Finally, there is the activation. In DUAL, there are two sources of activity: input nodes and goal nodes. Input nodes are the percepts; they emit a constant amount of activation for as long as the corresponding objects are part of the environment. Goal nodes are agents that have been identified as pragmatically important, and they emit a constant amount of activation for as long as they are on the goal list. □

We have now looked at the DUAL architecture, and seen the way in which its symbolic and connectionist aspects interact. But how does *AMBR* actually represent knowledge and reason by analogy? To answer these questions we will need to introduce three new types of agents — *concept* agents, *instance* agents and *hypothesis* agents. (AMBR also introduces four new symbolic processes, but we will start our review by looking at its agents.)

In AMBR, basic declarative knowledge is represented by concept and instance agents, while propositions are represented by small, inter-connected groups of agents called *coalitions*. Concept agents define *types* of objects and relations, while instance agents represent specific instances of objects, relations and situations. Each of these agents has a frame-like structure called a *micro-frame*; these micro-frames are part of the agents' symbolic components, and their slots are essentially pointers to other agents.

Coalitions also include an agent, known as the *coalition leader*, that 'represents' the coalition and the knowledge it expresses. These leaders will typically point to the essential elements of their coalition and the concepts that they derive from, or instantiate. If the coalition itself represents a concept, then the leader may also point to its prototypical instances.

And then we have situations — the 1994 version of AMBR used comprehensive lists of pointers to provide complete and rigid episode descriptions. While these centralised records of events simplified the mapping process, they were left out of subsequent versions due to their "psychological implausibility", and because doing so made it possible to investigate other phenomena including episode blending and intrusions from general knowledge.

In 1998, a more pliable approach to episode representation was adopted. Instead of providing the definitive account of an episode, the primary role of a situation agent is to stand for its spatio-temporal boundaries — the situation agent still points to the most salient features of an episode, however *it* is pointed to by all of the participating agents. Not only is this approach more psychologically plausible, but a decentralised representation makes it easier for AMBR to support context-dependent episodes and re-representation.

Finally, we have the hypothesis agents, or *hypotheses*, which represent potential correspondences between concept or instance agents. Unlike other mapping engines (such as ACME

[2]For brevity's sake, I shall use the terms *micro-agent*, *DUAL agent* and *agent* interchangeably, but only in this section.

and the SME, which enumerate all mappings that satisfy a simple, syntactic constraint), every hypothesis in AMBR must have at least one justification — be it semantic or structural. Semantic justifications indicate that two agents are semantically similar (e.g. that they are derived from a common super-class); structural justifications follow from their semantic counterparts and indicate that two hypotheses are consistent (e.g. the hypothesis that two relations correspond justifies the hypotheses that their arguments also correspond).

Having introduced AMBR's agents, we will now look at its mechanisms. Building, as it does, upon DUAL, spreading activation and marker-passing play an important role in AMBR; on top of these, however, AMBR introduces four more mechanisms — *structure-correspondence*, *constraint satisfaction*, *rating and promotion* and *skolemization*. The structure-correspondence, marker-passing and spreading activation mechanisms encourage structural, semantic and pragmatic correspondences, respectively; the constraint satisfaction mechanism ensures that consistent hypotheses are favoured over contradictory ones; the rating and promotion mechanisms decide which hypotheses go on to become winners; and the skolemization process implement a weaker form of transference.

These mechanisms focus on *local* correspondences; the constraint satisfaction mechanism ensures that these correspondences are *globally* consistent, and it does this by constructing a *constraint satisfaction network (CSN)*. This CSN is not unlike the network used by ACME, however there are several important differences: (1) AMBR constructs its CSN incrementally (i.e. *as* the hypotheses are created) and its 'solutions' can be read before the network settles; (2) the CSN is integrated into working memory, which gives it direct access to the system's goals and knowledge base, and which allows the retrieval and mapping processes to adjust the levels of activation in working memory; (3) the CSN only contains hypotheses that have a semantic or structural justification; and (4) AMBR's lack of rigid episode representations, and the incremental way in which it generates its CSN, means that the winning hypotheses may, in fact, come from a mixture of situations [25].

The winning hypotheses, in turn, are chosen by the rating and promotion mechanisms. In a system like ACME, the winners are simply the nodes with the highest asymptotic levels of activation; AMBR, however, uses a different criterion. AMBR's rating mechanism monitors the activation of competing hypotheses, and increases the rating of the most active hypothesis whilst decreasing the ratings of the others; the amount by which the ratings are changed is proportional to the difference in activation between the leader and its closest competitor.

When a hypothesis' rating drops below a *critical loser* threshold, it is removed from working memory; when a hypothesis' rating reaches an upper threshold, several things happen. First, one of the agents from the target situation checks to see if the leading hypothesis is consistent with the other mappings. If it is, then the leader is promoted to *winner*, and the ratings of its competitors are drastically reduced; if, however, the leader is incompatible with its neighbouring mappings then an inhibitory link is created between its arguments, its rating is reset, and the monitoring continues.

The last of AMBR's mechanisms is called skolemization. Essentially, skolemization is the process of creating specific propositions from general ones — that is, it is how AMBR applies general knowledge. Skolemization is triggered by one of two events: (1) when the activity of a general hypothesis exceeds a certain threshold, and (2) when the rating and promotion mechanism cannot find a clear leader using the specific propositions alone. □

In this section we looked at two models of analogy — the SME and AMBR. The original SME is one of the canonical models, and it embodies many of the assumptions of traditional theories: it defines analogy in terms of isomorphisms and syntactic rules alone, it operates over complete (and rigid) episode representations, it does not take semantic, pragmatic or contextual factors into account, and when it *is* combined with a model of analogue retrieval, the resulting system is sequential — i.e. the retrieval and mapping stages are largely independent.

Set against this, we have AMBR which *does* take structural, pragmatic and semantic constraints into account, which operates on decentralised and context-dependent representations of concepts and episodes, and which treats the retrieval, mapping and transferral stages as overlapping, and interacting, processes. In fact, AMBR is more than just a model of analogy — it parallels Cowan's model of working memory (see Section 2.2), and because the results of these stages can be accessed before its network settles, AMBR could be used as an 'online', or continually running, model of working memory.

While AMBR is both exciting and intuitive, it still falls short of the criteria given in Section 2.3; in particular, AMBR's ability to learn is quite limited — it is unable to identify (by itself) the salient or statistically significant features of an episode, and it is unable to construct new representations of episodes or concepts.

4. CONTEXT IN AI

For an online system to *create* an account of an episode, though, it needs more than just percepts — it also needs a way of discovering the scope, or extent, of the episode. This, in turn, requires information about the *types* of episodes that are likely to be encountered, and an explicit representation of these contexts. In the previous section we saw how context-*dependent* structures emerge from AMBR's use of activation and coalitions. In this penultimate section, we are going to look at the concept of context, and review, albeit briefly, how contexts are represented by agents and used to guide their behaviour.

At its highest level, context serves two purposes: (1) it allows us to explicitly represent all that is implicit about our environment, goals and assumptions; and (2) it allows us to constrain and qualify knowledge — that is, it makes knowledge and reasoning local. In the field of AI, context has been used to process natural language documents and database queries, and during knowledge acquisition (where it is used

to partition knowledge bases into smaller, locally consistent, modules) [5].

But what exactly does the term 'context' mean? In a general sense, context can refer to the temporal, spatial or conceptual 'locality' of an object or event. This is what Kokinov refers to as *external* context [16]; he also defines *internal* context as the agent's current mental state, which in our case, corresponds to the distribution of activation in long-term memory, and is shaped by perception, memory access and reasoning.

More pragmatically, context can also be defined as the subset of working memory that predicts, most succinctly, the assumptions and behaviours that an agent should adopt in order to maximise its measure of performance. This is consistent with Turner's definition of context as "any identifiable configuration of environmental, mission-related, and agent-related features that has predictive power for an agent's behaviour. The term situation is used to refer to the entire set of circumstances surrounding an agent, including the agent's own internal state." [28, p. 2]

With this pragmatic definition in mind, let us now look at some of the ways in which context, and context-dependent behaviour, can be represented. There are, of course, many such ways, but we will be focusing on representatives of two quite different approaches — one that uses logic, and one that uses micro-contexts to describe situations. In between these two approaches, lie techniques such as Context-based Reasoning (CxBR) and Contextual Graphs (CxG) [20].

(The CxBR model is essentially a state machine; contextual knowledge is asserted within discrete and mutually-exclusive states called *contexts*, and *sentinel rules* define the valid transitions. Contextual graphs, on the other hand, are directed acyclic graphs that consist of *action*, *contextual* and *recombination* nodes. Contexts are hierarchical, and are spanned by contextual–recombination node pairs; an agent's current context is determined by the presence of variables (represented by contextual nodes) and their values (that determine which arc, or branch, to follow). Unfortunately, neither of these approaches are flexible enough for an artifical memory system: in CxBR, contexts are monoliths that need to be defined explicitly, while CxGs focus more on the processes, or workflows, than the environment or the state of the agent.)

4.1 Representing Context in Logic

The first way in which we can represent context is through the use of first-order logic — an approach that allows us to represent the implicit aspects of a situation, and to constrain knowledge. But context in logic can also be used to *generalise* knowledge; as McCarthy noted in his 1993 paper on the subject, one of the goals of representing context in logic is to allow "simple axioms for common sense phenomena ... to be lifted to contexts involving fewer assumptions" [22].

In this paper, he argued that contexts should be treated as first class objects, and he described the basic relations and rules needed to define, and reason about them. The fundamental relation is $ist(c, p)$, which asserts that the proposition p is true in context c; in McCarthy's proposal, these relations are themselves defined within another context. He also defined the term $value(c, t)$, which returns the value of term t in context c; while the range of *value* is usually fixed, this term allows us to introduce context-dependent vocabularies.

In addition to these fundamental terms, however, this approach depends upon axioms that describe the contexts, and define their interrelationships; the latter axioms, or lifting rules, define how formulas can be lifted from one context to another. Through them, we can express context-specific knowledge, and move *between* contexts, but defining these rules can be a time-consuming task in itself. □

While this approach has its advantages, it also requires a great deal of hand-crafted rules, an external set of contexts, and some form of context recognition. Most importantly, from my perspective, it is basically incompatible with the activation-based models of working memory and analogy.

4.2 Context-Mediated Behaviour

The other approach that we are going to look at here is called *Context-Mediated Behaviour* (CMB) [28, 1]. It is being developed by Turner and his colleagues at the University of Maine, where it provides the context-management framework for Orca — an agent that is designed to control long-range autonomous underwater vehicles (AUVs); as such, it is more concerned with context-appropriate behaviour and computational efficiency than the ability to reason about contexts, or generalise knowledge.

The essence of their approach is this: (1) situations can be broken into one or more contexts; (2) contexts are represented by frame-like structures, called *contextual schemas* (or *c-schemas*), which are used to identify contexts and guide the agent's behaviour; and (3) that differential diagnosis is used to find the schemas that provide the 'best' account of a situation.

In CMB, *situations* represent the agent's view of the world, and include all of the environment's observable features; *contexts* correspond to recognisable and recurring subsets of these features; and *schemas* represent the agent's context-specific knowledge. (As these schemas focus on particular aspects of an agent's operating environment, several schemas are usually needed to provide a sufficiently detailed account of the larger, and more complicated situations; this merged schema represents the *current context*.)

Contextual schemas may be further divided into two parts. The *descriptive* part of a c-schema lists the agents and objects that might been seen in that context, together with an estimate of their probability and importance; it may also describe the features and characteristics of the environment itself. The *prescriptive* part of a c-schema is where context-sensitive values and behavioural parameters are set; it is also where events, and procedures for handling them, are defined.

Before an agent can *use* these schemas, however, they need to be retrieved from long-term memory; in Orca, this process is performed by the Embedded Context-Handling Object (ECHO). The algorithm that ECHO uses to select the c-

schemas is based upon work done in the area of medical diagnosis. Essentially, this process consists of four steps (see [1] for a detailed description of the algorithm):

1. *Find the contexts that predict the situation's features.* When Orca's long-term memory is probed with a subset of the situation's features, it returns the most specific c-schemas that fit the probe. These schemas, or *candidates*, are given a score based upon the number of features that are (a) predicted by the c-schema and present in the situation, (b) predicted but absent, and (c) not predicted but present.

2. *Cull the candidate set.* Candidates with scores that are much less than the highest score are removed.

3. *Evaluate the 'diagnoses'.* While one or more important features remain unaccounted for, (a) create a *competitor set* for the highest-scoring candidate (in Orca, two schemas are competitors if the features predicted by one are a subset of the other's); (b) find the winner — if the difference between the two highest scores exceeds a threshold, then the highest ranking schema wins; otherwise, Orca tries to find out more about the predicted-but-absent features; and (c) add the winner to the current context, and remove the features that it predicts from the set of 'symptoms'.

4. *Merge the winners into the current context.* This is where the schemas, and their potentially conflicting information, are merged into a coherent whole. □

While context-mediated behaviour has shown some promise, the approach itself is still incomplete — the details of *when* the current context should be calculated (and recalculated) are still being worked out, and the rules that determine *how* conflicting schemas should be merged is the subject of ongoing research. This, in turn, has implications for some of the criteria outlined in Section 2.3; in particular, the lack of an appropriate way to trigger a re-evaluation of the current context limits its ability to correctly characterise its environment and operate continuously. Finally, the question of how *new* schemas might be created is yet to be looked at.

There are also some differences between CMB and my vision of an artificial memory system. For example, CMB does not retrieve analogues; in a mission-critical setting, this is probably the correct thing to do, but this ability can be quite useful in other domains. Nor does it support the cognitive process of episode blending; while CMB *does* create its representations of the current context by merging several smaller schemas, these schemas usually describe independent aspects of the situation. Episode blending, on the other hand, merges episodes that are superficially or structurally *similar* [14] — a precursor, perhaps, of the ability to make generalisations.

And finally, the fact that CMB only recalculates its current contexts in response to an external trigger makes it much coarser in time. Again, this might be appropriate for a mission-critical system, but it also make it less sensitive and responsive to changes in the environment. (Indeed, a continuous approach that uses spreading activation might even eliminate the need for an external trigger.)

In spite of these differences and limitations, CMB has shown that an artificial memory system — one that is separate from the reasoning system, and presents it with a context-dependent knowledge base — is, indeed, feasible. Furthermore, the work being done on schema merging may also be relevant to my project.

5. CONCLUSION

In this paper, we have looked at cognitive models of working memory and analogy, and we have touched on some of the ways in which context, and context-dependent knowledge have been represented in AI. While most of this work dates back to the early to mid-1990s, the cognitive models of human memory and analogy-making have not received a lot of attention from the AI community; even less work has been done to combine these models, or to extend their abilities to learn. The purpose of this review, then, is to encourage such an attempt.

Cowan's Embedded-processes model of working memory provides an empirical account of human memory and attention — both voluntary and involuntary. It does not, however, explain how we *acquire* these memories in the first place, nor does it specify which measures of similarity are used to spread the activation; and while he notes that awareness allows us to "incorporate contextual constraints that would not automatically be taken into account" [8, p. 176], he does not actually go on to define context.

Kokinov's Associative Memory-Based Reasoning (AMBR) is a hybrid, activation-based model of associative memory that is also able to retrieve analogues, and to discover correspondences between them. It does not, however, attempt to model our attentional system — unlike Cowan's model, AMBR includes neither a central executive nor the concept of habituation. AMBR also lacks the ability to learn, and to acquire explicit representations of concepts *or* contexts.

And finally, we looked at Turner's model of Context-Mediated Behaviour (CMB). While it bears some resemblance to AMBR, it does differ from the other models in several ways: (1) it is *not* a cognitive model, but a way of providing artificial agents with context-sensitive knowledge; (2) unlike AMBR (which uses thresholds and a continual spreading of activation to determine the contents of its working memory), CMB only re-evaluates its current context after a significant change has occurred; (3) it represents contexts, or *types* of situations instead of episodes; and (4) it represents knowledge much more coarsely and rigidly than AMBR.

While CMB has shown that the *concept* of artificial working memory is, indeed, feasible, a memory system that is based upon the cognitive models of Kokinov and Cowan will offer much more flexibility — it will serve as a context-sensitive knowledge base, but it will also be able to retrieve knowledge based upon literal and structural measures of similarity. If we can add to this, the ability to learn — i.e. to discover statistically significant associations and sub-graphs — it may even be able to acquire representations of *classes* of episodes, and through this, the ability to identify contexts in an online environment.

6. REFERENCES

[1] R. P. Arritt and R. M. Turner. Situation Assessment for Autonomous Underwater Vehicles Using A Priori Contextual Knowledge. In *Proceedings of the 13th International Symposium on Unmanned Untethered Submersible Technology*, Durham, NH, 2003.

[2] A. D. Baddeley and G. J. Hitch. *Working Memory*, volume 8 of *The Psychology of Learning and Motivation: Advances in Research and Theory*, pages 47–89. Academic Press, New York, 1974.

[3] A. D. Baddeley and R. H. Logie. Working Memory: The Multiple-Component Model. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, chapter 2, pages 28–61. Cambridge University Press, 1999.

[4] J. G. Benjafield. *Cognition*. Prentice-Hall, 1992.

[5] P. Brézillon. Context in Problem Solving: A Survey. *The Knowledge Engineering Review*, 14(1):1–34, 1999.

[6] D. E. Broadbent. *Perception & Communication*. Pergamon, New York, 1958.

[7] D. E. Broadbent. The Maltese Cross: A New Simplistic Model for Memory. *The Behavioral & Brain Sciences*, 7:55–94, 1984.

[8] N. Cowan. Evolving Conceptions of Memory Storage, Selective Attention, and Their Mutual Constraints Within the Human Information-Processing System. *Psychological Bulletin*, 104(2):163–191, 1988.

[9] N. Cowan. An Embedded-Processes Model of Working Memory. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, chapter 3, pages 62–101. Cambridge University Press, 1999.

[10] K. A. Ericsson and P. F. Delaney. Long-Term Working Memory as an Alternative to Capacity Models of Working Memory in Everyday Skilled Performance. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, chapter 8, pages 257–297. Cambridge University Press, 1999.

[11] B. Falkenhainer, K. D. Forbus, and D. Gentner. The Structure-Mapping Engine. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, August 1986.

[12] B. Falkenhainer, K. D. Forbus, and D. Gentner. The Structure-Mapping Engine: Algorithms and Examples. *Artificial Intelligence*, 41:1–63, 1989.

[13] D. Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2):155–170, 1983.

[14] M. Grinberg and B. Kokinov. Analogy-Based Episode Blending in AMBR. In *Constructive Memory*, pages 157–167. NBU Press, 2003.

[15] K. J. Holyoak and P. Thagard. Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, 13:295 – 355, 1989.

[16] B. Kokinov. A Dynamic Theory of Implicit Context. In *Proceedings of the 2nd European Conference on Cognitive Science*, pages 252–255, April 1997.

[17] B. N. Kokinov. A Hybrid Model of Reasoning by Analogy. In *Analogical Connections*, volume 2 of *Advances in Connectionist and Neural Computation Theory*, pages 247–320. 1994.

[18] B. N. Kokinov and A. A. Petrov. Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In *The Analogical Mind: Perspectives from Cognitive Science*, chapter 2, pages 59–124. MIT Press, Cambridge, MA, 2001.

[19] C. Lebiere and J. R. Anderson. A Connectionist Implementation of the ACT-R Production System. In *Proceedings of the Fifteenth Conference of the Cognitive Science Society*, pages 635–640, Hillsdale, NJ, 1993. Erlbaum.

[20] P. Lorins, P. Brézillon, and A. Gonzalez. Context-based Decision Making: Comparison of CxBR and CxGs Approaches. In *Proceedings of the IFIP International Conference on Decision Support Systems*, pages 115–124, 2004.

[21] J. McCarthy. First Order Theories of Individual Concepts and Propositions. In Donald Michie, editor, *Machine Intelligence 9*. Ellis Horwood, 1979.

[22] J. McCarthy. Notes on Formalizing Context. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 555–560, 1993.

[23] A. Miyake and P. Shah, editors. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press, 1999.

[24] D. A. Norman. Toward a Theory of Memory and Attention. *Psychological Review*, 75(6):522–536, 1968.

[25] A. A. Petrov. *A Dynamic Emergent Computational Model of Analogy-Making Based on Decentralized Representations*. PhD thesis, Central and Eastern European Center for Cognitive Science, New Bulgarian University, July 1998.

[26] A. A. Petrov and B. N. Kokinov. Processing Symbols at Variable Speed in DUAL: Connectionist Activation as Power Supply. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 846–851, San Francisco, CA, 1999. Morgan Kaufman.

[27] P. Thagard, K. J. Holyoak, G. Nelson, and D. Gochfeld. Analog Retrieval by Constraint Satisfaction. *Artificial Intelligence*, 46(3):259–310, 1990.

[28] R. M. Turner. Context-Mediated Behaviour for Intelligent Agents. *International Journal of Human-Computer Studies*, 48(3):307–330, 1998.

[29] P. H. Winston. Learning and Reasoning by Analogy. *Communications of the ACM*, 23(12):689–703, 1980.