An Overview of: Logic, Self-awareness and Self-improvement: the Metacognitive Loop and the Problem of Brittleness [1]

CMSC 828D
Fall 2012
Writing Assignment 1
Michael Maynord

The field of AI has had many successes over the decades, in many instances machines consistently perform at levels beyond the capacity of any human, the task of playing chess is a good example. In many other instances, what humans find trivial, machines find incredibly difficult. One important such instance is dealing gracefully with unforeseen situations; that is, not breaking because of some minor perturbation in the problem scenario. This problem in AI is referred to as the "brittleness" problem – virtually all contemporary AI systems will break if presented with scenarios which they were not explicitly designed to cope with. An example of brittleness is an autonomous vehicle running into a chain link fence which its sensors are unable to detect, and instead of backing up and finding an alternative route, simply spinning its wheels[1]. Brittle systems lack what we refer to as "common-sense". The question, then, is how to impart common-sense onto machines?

Common-sense involves having expectations about how the world is, or how it should be when considered with respect to certain goals, and having the capacity to detect when these expectations are not in accord with the observed state of the world. When observation does not match expectation, some rectifying action must be initiated - either change expectations to match the world, or change the world to match expectations. Intuitively this is what humans are doing when they apply common-sense – rectifying observation/expectation mismatches.

Three Difficulties

Three difficulties which confound working with expectation/observation discrepancies are: slippage, KR mismatch, and contradiction[1].

Slippage occurs with unchanging beliefs about a world which is subject to change through time. For example, if yesterday robot A formed the belief that cat B is at point C, and still believes that today though today the cat is now at point D, slippage has occurred(The applicability of the belief has slipped into the past).

KR mismatch is a problem with multiple representations(multiple syntax's) for describing the world – how should different representations be dealt with?(Two pictures of a single tree but from a different angle present different representations, but representations of the same object.)

Contradictions arise from belief bases which evolve with observations of the world. Consistency in a belief base which is dynamically constructed from observations of a complex and dynamic world is prohibitively difficult to maintain, and contradictions are inevitable in any system powerful enough to reflect our world.(The weather report says it's 50 degrees Fahrenheit, however you see through the window that it is snowing, this is a contradiction that must be dealt with gracefully, a proper reaction might be to conclude that the weather report is erroneous.)

One approach for reasoning in the face of slippage, KR mismatch, and contradictions is active logic[1]. Active logic is a kind of para-consistent logic which takes time into consideration. It has an evolving time frame in which beliefs which are subject to change through time are forced to actively perpetuate themselves, or to evolve. This directly confronts the issue of slippage.

A feature of active logic is that it does not induce all the implications of its present belief set at once, rather it induces stepwise along with explicit time steps. Thus, multiple beliefs which together may approximate the world, but entail contradictions, do not break the system, as contradictions are dealt with when(in the explicit time steps of active logic) and if they are explicitly induced. When contradictions arise, active logic, rather than breaking, is simply careful when using the contradictands in any further inductions, until one of the contradictands has been proved true, and the other false. This feature helps confront both the issue of contradictions and KR mismatch.

Meta-Cognition

An approach to confronting the problem of brittleness is to introduce into AI systems some degree of meta-cognition, that is, give AI systems the capacity to think about their own thinking process. In the instance of the autonomous vehicle, the object of first level cognition is the task of getting from point A to point B, the task of meta-cognition is to direct the thought process governing this task. Whereas first level cognition was unable to deal with the fence, because it was undetectable by sensors, meta-cognition should be able to detect the presence of a problem and direct the first level cognition to pursue an alternative means to its goal.

One meta-cognitive approach which leverages the advantages of active logic is the Meta-Cognitive Loop(MCL)[1]. MCL combines “neat” and “scruffy” approaches to AI.(The “neat” school of thought in AI says approaches should be elegant, the “scruffy” school of thought says the most effective methods come from mashing a whole bunch of sub-methods together each of which empirically work for their given sub-task.) In MCL the cognitive level is scruffy, and the meta-cognitive level is neat. The cognitive level consists of many trainable modules suitable for various tasks, as well as trainer modules to train these trainable modules. The meta-cognitive level takes an ordered approach to determine which modules should be applied to which tasks, and when and how modules should be retrained.

MCL functions by cycling through three stages: Note, Assess, and Guide. In the Note stage, MCL considers percepts to determine if there is a discrepancy between observation and expectation. In the assess stage, the type of anomaly is assessed with respect to the type of solutions that could be applied, and in the guide stage the system is guided through the response to the anomaly.

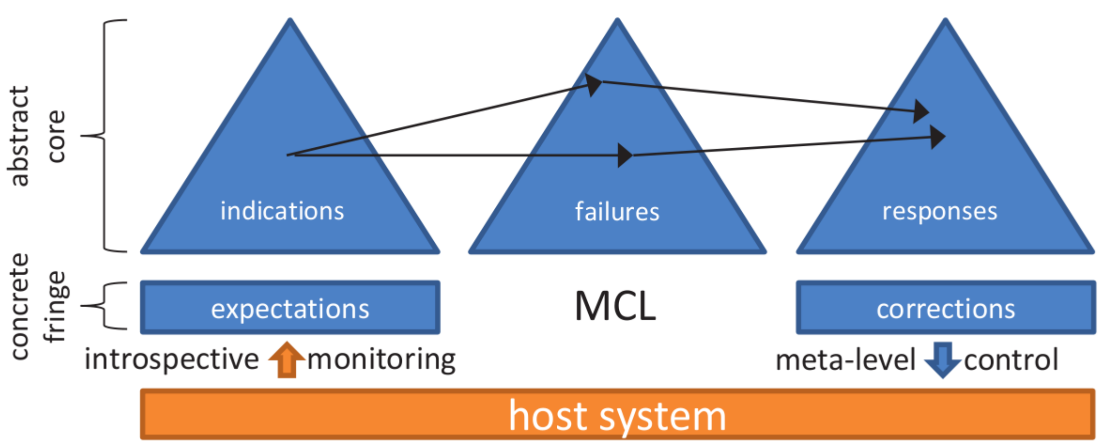

Figure 1: Meta-Cognitive Loop[2]

A different paper, Toward an Integrated Metacognitive Architecture[2], by Michael T. Cox, Tim Oates and Don Perlis, provides figure 1 and gives a brief description. Essentially, blue corresponds to meta-cognitive, orange corresponds to cognitive, “indications” corresponds to noting an anomaly, “failures” corresponds to assessing it, and “responses” corresponds to guiding. The three abstract cores are instantiated as Bayes Nets, with peripheral nodes corresponding to domain specific conditions, and more central nodes corresponding to conditions which are more invariant across domains. It is in part these relatively domain independent nodes of the Bayes Net that gives MCL versatility across domains.

Successes

There are at least three examples where MCL has been applied and shown to be of benefit[1].

The first example is applying MCL to a reinforcement learner in a dynamic environment. When a standard reinforcement learner has learned a given approach in a relatively static environment, and the environment changes drastically and suddenly, the learner takes a fairly long time to learn a new strategy suitable to the changed environment. However, MCL can be applied in such a way that when the environment changes suddenly, it is recognized that the unusual and sudden poor performance of the learner is likely a result of a change in the environment, and a new strategy should be learned post-haste, and the old strategy thrown out.

A second example is MCL enhanced navigation. The scenario is a neural net learning to navigate in an environment in which it can collide with obstacles. When MCL is applied, it is able to recognize collision events and apply targeted alterations to the neural nets; alterations such as retraining with a different initial weight set, or retraining using only specific inputs. With MCL, the neural net reaches a proficient level faster than without.

A third example is a human-computer dialog application. Applying MCL allows the application to keep track of its understanding of language. In an example given[1], the user states “Send the Boston train to New York”, a train is then dispatched and the user says “No, send the *Boston* train to New York”, it is then reasoned that the wrong train was selected and instead of dispatching the same train a second time, a different train is sent. In another example, the user says to send the metro, and the system is able to recognize 'metro' as an unrecognized word and ask for its meaning(metroliner). The ability to reason with meta-language like this can be of significant help in communication.

Conclusion

The brittleness problem is the problem of machines lacking common-sense. In order for machines to gain common-sense, they should gain the ability to construct a structured expectation system and the ability to compare expectations with observations. Difficulties which arise are slippage, KR mismatch, and contradictions. Active logic moves to confront these issues and is incorporated into MCL. MCL cycles through note, assess, and guide phases, and has had success in applications such as reinforcement learning, navigation, and human-computer dialogue.

Sources:

[1] Michael L. Anderson and Donald R. Perlis (2005), Logic, Self-awareness and Self- improvement: the Metacognitive Loop and the Problem of Brittleness ,J Logic Computation (2005) 15(1): 21-40
http://logcom.oupjournals.org/cgi/content/abstract/15/1/21?ijkey=01seTUWa5vvtU&keytype=ref

[2] Cox, M. T., Oates, T., & Perlis, D. (2011). Toward an integrated metacognitive architecture. In P. Langley (Ed.), *Advances in Cognitive Systems, papers from the 2011 AAAI Symposium* (pp. 74-81). Technical Report FS-11-01. Menlo Park, CA: AAAI Press.
https://skydrive.live.com/?cid=fdd01a401da67d10&id=FDD01A401DA67D10!140