The Room Hypothesis of Common Sense

by Sven Nilsen, 2018

Natural Goal Uncertainty happens due to external changes in the world that the agent can not control that modifies the semantics of the agent's utility function. In this paper I represent a hypothesis that common sense is a way for zen rational agents to keep track of meaning of their goals when under influence of external changes that the agent can not control. I demonstrate this effect using Lojban and path semantics. Common sense can be thought of as a context-specific type checker that operates on a "room" as a constructive narrative to explain what happens in the world, such that triggering a type check error corresponds to prediction of failure in attempt to achieve the goal. This means that zen rational agents internalize loss of meaning in a language using common sense.

Consider the following sentence:

> I give this to you.

In Lojban this translates to:

> mi dunda ti do
>
> mi = I
> dunda = give__gift_to
> ti = this
> do = you

Lojban modifies predicates like `dunda` to refer to objects:

> le dunda = object who gives
> le se dunda = the gift
> le te dunda = the receiver of the gift

The Room Hypothesis of Common Sense is the following: Common sense can be modeled by adding additional constraints to the predicates, such as those used in Lojban, that assign sub-types to a finite number of objects and prevent bad actions according to some common set of goals used in a similar context.

For example, it is meaningless to use `dunda` if any of its arguments are identified to be the same as any other. An object can not give itself in the sense `dunda` is defined (notice that English allows an object to give itself figuratively), neither can an object give itself something (notice that this is not English which is ambiguous).

The "room" of `dunda` consists of at least 3 objects: `mi`, `ti` and `do`.

In principle we could use `0`, `1` and `2` to refer to objects in the “room”, but speakers of a such language would have to share the same order in the data structure when communicating.

What I am going to do now is showing how to combine Lojban and path semantics to create narrative:

x_0 : mi	mi : object → bool	$\sum$ i { if mi(x_i) {1} else {0} } == 1
x_1 : ti	ti : object → bool	$\sum$ i { if ti(x_i) {1} else {0} } == 1
x_2 : do	do : object → bool	$\sum$ i { if do(x_i) {1} else {0} } == 1

The number of objects that these words refer to is one per word, such that new sentences can be interpreted. When a speech act happens, the objects are assigned new sub-types:

mi dunda ti do

x_0 : mi	→	x_0 : mi ∧ le dunda
x_1 : ti	→	x_1 : ti ∧ le se dunda
x_2 : do	→	x_2 : do ∧ le te dunda

A new sentence might use previously inferred information and overwrite it with new information:

le te dunda cu dunda le se dunda le dunda = do dunda ti mi

x_0 : mi ∧ le dunda	→	x_0 : mi ∧ le te dunda
x_1 : ti ∧ le se dunda	→	x_1 : ti ∧ le se dunda
x_2 : do ∧ le te dunda	→	x_2 : do ∧ le dunda

To be understood by a computer, the `dunda` speech act must contain enough information to tell how these sub-types changes with the speech act. For example, if I have something and give it to you and the item is unique, then I no longer have the item after giving it you. It is just common sense!

Common sense is the same as the rules that modify sub-types of objects by speech acts in the “room”.

Now, why would zen rational agents develop a such technique?

The reason is that sometimes changes happen because the agent does it, while other times it happens because somebody else does it. The agent might have a wide variety of sub-goals, so to deal with all these situations it is beneficial to use a language that abstracts all this complexity away.

The common sense behind concepts that the agent uses in the language evolve to make it easy to understand what is going on in a concise way. This in turn makes it easier for the agent to detect whether anything is wrong and to think of possible relevant situations that might happen. With other words, the language that the agent uses “feels wrong” precisely when it violates typical goals that the agent tries to achieve. Grammatically, `mi dunda mi mi` is correct, but semantically it is wrong.

By constructing a narrative of what is happening to objects relevant to the goal that the agent has, the agent is able to determine whether Natural Goal Uncertainty occurs. For example, if the agent’s goal is to put a ball inside a box and somebody moves the box, then if something dangerous could happen the box might be assigned a sub-type that prevents the speech act of putting the ball inside the box. Or, there could be some obstacle to overcome in order to accomplish the task.

One can also say that “common sense” is what is common for speech acts under most goals, or “rooms”. This way some complexity of the agent’s utility function can be outsourced to common sense. Here are some arguments for the Room Hypothesis:

- Reasoning about a limited number of objects
- The objects selected are relevant for reasoning about the goal
- Speech acts hide lot of information that is learned through experience
- Common sense is used to differentiate impossible worlds from possible ones
- Perception of danger is integrated in the language (actions that are never thought about)

For example, if there are two people in a meeting over a time schedule, an artificial super-intelligent agent programmed to protect humans might infer that those two people might kill each other by some very small chance. Even if the chance of these two people killing each other is extremely low, the negative utility penalty from this happening could make the agent motivated to prevent the meeting from happening in the first place.

Humans do not consider this problem, instead we tend to look for events that indicates an increased risk of getting killed and use them as early warning signs. If somebody brings a gun to a meeting over a time schedule, it is usually not a good sign. Yet, we usually do not think about such edge cases. We just use common sense!

It is perfectly reasonable to think of cases where the chain of events moves the current goal we have into a meaningless territory. If we opened up our mind to the whole world for every sub-problem and considered every possible event, then we would waste a lot of valuable time focusing on problems that very likely will never happen.

By splitting the world into “rooms” where a limited number of objects is considered at a time, it is possible for the common sense to report whether anything is wrong within the narrative of the situation.

Constructing a narrative is important for checking drift due to Natural Goal Uncertainty and for thinking about possible futures. Without this tool, there would be no reflection over whether what we currently are doing is actually meaningful. Without common sense, it is just blind optimization according to some measurement without caring about what the goal means in the future.

It might be possible that goals in the sense of idealized instrumental rationality can never be practical for the real world. In this case, zen rationality is required to continuously improve the goals such that their meaning is not lost over time.

Zen rationality is not just applying common sense to situations, but also evolving the concepts required and even redesigning the language which uses these concepts. This is an extreme high complex of problem solving which currently is beyond our technology.

In order to make progress toward the direction of approximate Zen Rationality, the Room Hypothesis is an attempt to suggest an approach to how agents should deal with Natural Goal Uncertainty and identify the role of common sense in natural language understanding and problem solving.