A Pragmatic Path Toward Endowing Virtually-Embodied AIs with Human-Level Linguistic Capability

Ben Goertzel

Novamente LLC, 1405 Bernerd Place, Rockville MD 20851, USA

Abstract— **Current work is described wherein simplified versions of the Novamente Cognition Engine (NCE) are being used to control virtual agents in virtual worlds such as game engines and Second Life. In this context, an IRC (imitation-reinforcement-correction) methodology is being used to teach the agents various behaviors, including simple tricks and communicative acts. Here we describe how this work may potentially be exploited and extended to yield a pathway toward giving the NCE robust, ultimately human-level natural language conversation capability. The pathway starts via using the current system to instruct NCE-controlled agents in semiosis and gestural communication; and then continues via integration of a particular sort of hybrid rule-based/statistical NLP system (which is currently partially complete) into the NCE-based virtual agent system, in such a way as to allow experiential adaptation of the rules underlying the NLP system, in a manner that builds on the agent's knowledge of semiosis and gesture.**

I. Introduction

Artificial intelligence technology currently does a lot of interesting and valuable things, but, it lacks the broad-based, creative general intelligence possessed by humans, a fact that has led Ray Kurzweil [1] to introduce an explicit distinction between purpose-specific "narrow-AI" and more flexible, autonomous "strong AI" (his term) or "Artificial General Intelligence" (AGI, our preferred term) [2]. A growing and increasingly vocal minority of the AI research community now believes that powerful artifical general intelligence (AGI) at the human level and beyond is reasonably likely to occur within the present century [3]. A smaller minority believes that AGI at the human level or beyond could be achieved more rapidly, even within the next 1-2 decades, if a sufficiently concerted effort were exerted in connection with the right set of ideas. We belong to the latter, smaller, more ambitious and optimistic minority.

Furthermore, we do not believe there is any single "golden path" to powerful AGI: rather we suggest that, due to the broad and heterogeneous nature of intelligence itself, there are bound to be multiple pathways, which will have different strengths and weaknesses at different stages in their development. For instance, closely neuromorphic architectures (such as [4, 5]; see [6] for a survey) will most likely have early strengths related to perception and action processing, whereas architectures based on explicit logic (such as [7,8]; see also [6]) will most likely have early strengths in language and mathematics.

Our own venture in practical AGI design and engineering, the Novamente Cognition Engine or NCE [9,10,11], is integrative in nature, relying on a weighted-labeled-hypergraph knowledge representation that fuses connectionist and logicist ideas, and incorporating a number of probabilistic learning mechanisms including PLN probabilistic logical inference [12, 13] and MOSES probabilistic evolutionary learning [14, 15]. Components of the NCE have been utilized in a variety of commercial applications, mainly involving data mining and language processing; and a simplified version of the NCE is now being used to control embodied virtual agents in virtual worlds such as game engines and Second Life [16]. We have argued elsewhere [17] that virtual-world embodiment provides a powerful medium for the experiential and instructional education of AGI systems.

While there are many potential paths to powerful AGI, the one we will discuss here is specific and language-centric. It seems clear that, once an AGI system has been created that is capable of reasonably robust and general NL conversation, then a massive acceleration of AGI progress will follow. Once we can really, flexibly talk to an AGI, we will be able to teach it all sorts of things, and we will be able to very sensitively gauge the impact of various internal changes on the AGI's general intelligence. So one interesting class of pathways to AGI consist of those that focus on language comprehension and production at a fairly early stage. Note that this is different than focusing initially on creating AGIs that can pass the Turing Test [18]. The ability to fool a judge that one is human is a specific cognitive ability that is not necessarily implied by a useful and robust NL capability. We are not particularly interested in creating AGIs that are effective "impersonators," but rather effective communicators, learners and creators.

Suppose one decides that creating a passably generally-intelligent English conversationalist is a useful goal for an AGI project. What then is the right approach in terms of AGI architecture, embodiment, knowledge representation, learning, instructional methodology, and so forth? There are many different possibilities on all these fronts: chat-bot-style conversation versus robotic or virtually-embodied conversation; a human-created NLP framework versus language learning via a generalized learning faculty; etc. Here we describe a specific pathway toward robust NL

Ben Goertzel (e-mail: ben@goertzel.org) works at Novamente LLC, Rockville, Maryland, USA

conversation, which was conceived with the NCE in mind, but also has potential applicability beyond this particular AGI architecture. The pathway we describe here combines normally-disparate aspects: It is not initially focused on language but on embodied sensorimotor learning; yet it also involves the creation and ensuing adaptation of a human-coded NLP functionality.

In brief, we suggest that the best route to creating a conversationally-capable AI is to begin with a virtually-embodied AI with a flexible learning capability, and instruct this AGI in the ways of semiosis and gesture. Once this foundation has been learned, it then makes sense to interact with the AGI linguistically. Potentially, this may be done via de-novo language learning; or, in what we suggest may be the most efficacious route, it may be done via supplying the AGI with a fully-featured but internally-very-flexible NLP subsystem, which may then be adapted based on the system's experience, which critically include semiotic and gestural interactions. The main content of this paper is a detailed discussion (insofar as space permits) of this multi-stage route to achieving advanced conversational ability in virtually-embodied AGI's. Before presenting this discussion, however, we review some of our current work on teaching virtually-embodied NCE-based agents, which forms the early stages of the proposed pathway to advanced NLP.

II. THE NOVAMENTE COGNITION ENGINE

One may decompose the overall task of creating a powerful AGI system into four aspects (which of course are not entirely distinct, but still are usefully distinguished):

-- 1. Cognitive architecture (the overall design of an AGI system: what parts does it have, how do they connect to each other)

-- 2. Knowledge representation (how does the system internally store declarative, procedural and episodic knowledge; and how does it create its own representation for knowledge of these sorts in new domains it encounters)

-- 3. Learning (how does it learn new knowledge of the types mentioned above; and how does it learn how to learn, and so on)

-- 4. Teaching methodology (how is it coupled with other systems so as to enable it to gain new knowledge about itself, the world and others)

We now briefly review how these four aspects are handled in the NCE. For a more in-depth discussion of the NCE the reader is referred to [9-11].

-- 1. The NCE's high-level cognitive architecture is motivated by human cognitive science and is roughly analogous to Stan Franklin's LIDA architecture [19]. It consists of a division into a number of interconnected functional units corresponding to different specialized capabilities such as perception, motor control and language, and also an "attentional focus" unit corresponding to intensive integrative processing. A diagrammatic depiction is given in [9].

-- 2. Within each functional unit, knowledge representation is enabled via an AtomTable software object that contains nodes and links (collectively called Atoms) of various types representing declarative, procedural and episodic knowledge both symbolically and subsymbolically. Each unit also contains a collection of MindAgent objects implementing cognitive, perception or action processes that act on this AtomTable, and/or interact with the outside world.

-- 3. In addition to a number of specialized learning algorithms associated with particular functional units, the NCE is endowed with two powerful learning mechanisms embedded in MindAgents: the MOSES probabilistic-program-evolution module (based on [14,15]), and the Probabilistic Logic Networks module for probabilistic logical inference [12,13]. These are used both to learn procedural and declarative knowledge, and to regulate the attention of the MindAgents as they shift from one focus to another, using an economic attention-allocation mechanism that leads to subtle nonlinear dynamics and associated emergent complexity including spontaneous creative emergence of new concepts, plans, procedures, etc.

-- 4. Teaching methodology is the main subject of this paper. We advocate a virtually-embodied approach which integrates linguistic with nonlinguistic instruction, and also autonomous learning via spontaneous exploration of the virtual world.

III. VIRTUALLY EMBODIED LEARNING WITH THE NCE

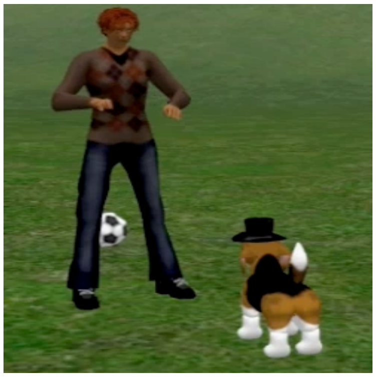

Fig. 1. Screenshot of a virtual animal in Second Life, controlled by the NCE-based AGI architecture described in this section.

In this section we briefly describe our current work using a simplified version of the Novamente Cognition Engine (the so-called "Virtual Animal Brain" or VAB) to control virtual animals in the Second Life virtual world. Figure 1 above shows an example virtual animal controlled by the VAB, interacting with a human-controlled avatar in the context of learning to play soccer. Figure 2 gives a high-level architecture diagram for the VAB, which is a simplification of the overall NCE architecture as diagrammed in [9].

The capabilities of the VAB-controlled virtual animals, in their current form, include

-- Spontaneous exploration of the environment

-- Automated enactment of a set of simple predefined behaviors

-- Flexible trainability: i.e., (less efficient) learning of behaviors invented by teachers on the fly

-- Communication with the animals, for training of new behaviors and a few additional purposes, occurs in a special subset of English called ACL (Animal Command Language)

-- Individuality: each animal has its own distinct personality

-- Spontaneous learning of new behaviors, without need for explicit training

Capabilities intended to be added in future VAB versions include

-- Recognition of novel categories of objects, and integration of object recognition into learning

-- Generalization based on prior learning, so as to be able to transfer old tricks to new contexts

-- Use of computational linguistics to achieve a more flexible conversational facility (this will be discussed extensively in a later section)

The VAB architecture is not particular to Second Life, but has been guided somewhat by the particular limitations of Second Life. In particular, Second Life does not conveniently lend itself to highly detailed perceptual and motoric interaction, so we have not dealt with issues related to these in the current version of the VAB. However, we have dealt with some of these issues in a prior version of the VAB, which was connected to the AGISim framework, a wrapper for the open-source game engine CrystalSpace [20].

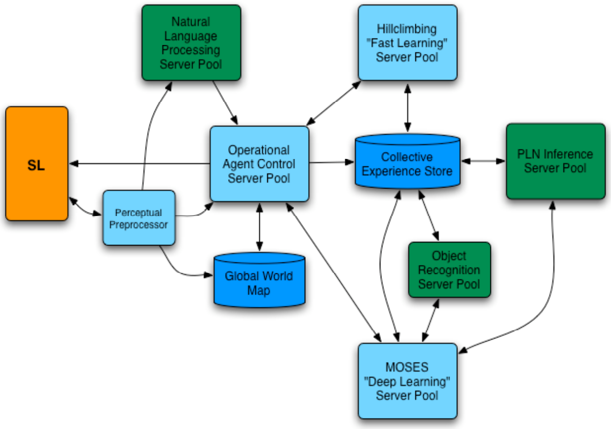
Natural Language Processing Server Pool
Hillclimbing "Fast Learning" Server Pool
SL
Operational Agent Control Server Pool
Collective Experience Store
PLN Inference Server Pool
Perceptual Preprocessor
Global World Map
Object Recognition Server Pool
MOSES "Deep Learning" Server Pool

Fig. 2. High-level diagram depicting VAB software architecture. The NLP, object recognition and PLN components are missing from the architecture that will initially be commercially deployed but are present in Novamente LLC's internal research codebase.

Instruction of VAB-controlled agents takes place according to a methodology we call IRC learning and is described in detail in [16], involving three interacting aspects:

-- *Imitative* learning: The teacher acts out a behavior, showing the student by example what he wants the student to do

-- *Reinforcement* learning: The student tries to do the behavior himself, and the teacher gives him feedback on how well he did

-- *Corrective* learning: As the student attempts the behavior, the teacher actively corrects (i.e. changes) the student's actions, guiding him toward correct performance

The combination of these three sorts of instruction appears to us critical, for learning of complex embodied behaviors and also, further along, for language learning. Current experimentation with the IRC methodology has been interesting and successful, resulting in a framework allowing human-controlled avatars to teach VAB-controlled agents a variety of behaviors such as fetching objects, delivering objects, going to desired locations, doing dances, and so forth.

The VAB work has been described in more depth in [16] and we will not repeat that discussion here; our present goal is, rather, to explain how one may potentially use an expanded version of it as a platform for creating NCE-controlled virtual agents with robust conversational capability. The path that we see from the VAB to robust English conversation has a series of stages, which we will now describe.

IV. TEACHING SEMIOSIS

The foundation of communication is semiosis. So before turning to the instruction of gesture or verbal language, we will treat the instruction of semiosis in itself, in a few relatively simple examples intended to make the principles clear. We will structure our discussion of semiotic learning according to Charles Sanders Peirce's theory of semiosis [21], in which there are three basic types of signs: icons, indices and symbols.

In Peirce's ontology of semiosis, an icon is a sign that physically resembles what it stands for. Representational pictures, for example, are icons because they look like the thing they represent. Onomatopoeic words are icons, as they sound like the object or fact they signify. The iconicity of an icon need not be immediate to appreciate. The fact that "kirikiriki" is iconic for a rooster's crow is not obvious to English-speakers yet it is to many Spanish-speakers; and the the converse is true for "cock-a-doodle-doo."

Next, an index is a sign whose occurrence probabilistically implies the occurrence of some other event or object (for reasons other than the habitual usage of the sign in connection with the event or object among some community of communicating agents). The index can be the cause of the signified thing, or its consequence, or merely be correlated to it. For example, a smile on your face is an index of your happy state of mind. Loud music and the sound of many people moving and talking in a room is an index for a party in the room. On the whole, more contextual background knowledge is required to appreciate an index than an icon.

Finally, any sign that is not an icon or index is a symbol. More explicitly, one may say that a symbol is a sign whose relation to the signified thing is conventional or arbitrary. For instance, the stop sign is a symbol for the imperative to stop; the word "dog" is a symbol for the concept it refers to.

The distinction between the various types of signs is not always obvious, and some signs may have multiple aspects.

For instance, the thumbs-up gesture is a symbol for positive emotion or encouragement. It is not an index -- unlike a smile which is an index for happiness because smiling is intrinsically biologically tied to happiness, there is no intrinsic connection between the thumbs-up signal and positive emotion or encouragement. On the other hand, one might argue that the thumbs-up signal is very weakly iconic , in that its up-ness resembles the subjective up-ness of a positive emotion (note that in English an idiom for happiness is "feeling up").

Teaching an embodied virtual agent to recognize simple icons is a relatively straightforward learning task. For instance, suppose one wanted to teach an agent that in order to get the teacher to give it a certain type of object, it should go to a box full of pictures and select a picture of an object of that type, and bring it to the teacher. One way this may occur in a NCE-controlled agent is for the agent to learn a rule of the following form (the semantics of the notation will be defined in following paragraphs)[1]

```
Implication
  AND
    Context
      Visual
      Similarity $X $Y
  PredictiveImplication
    SequentialAnd
      Execution goto box
      Execution grab $X
      Execution goto teacher
    Evaluation give me teacher $Y
```

While not a trivial learning problem, this is straightforward to an NCE-controlled agent that is primed to consider visual similarities as significant (i.e. is primed to consider the visual-appearance context within its search for patterns in its experience).

The notation in the above and following examples will hopefully be largely transparent to the reader, and has been defined formally and discussed extensively in [9]. Called "PLN notation" it is a textual notation used to depict nodes and links that exist in the AtomTable knowledge store of a NCE-based AGI system. Here we will define the notation only concisely and semi-formally, due to space considerations and to avoid redundancy with prior publications.

-- Indentation is used to signify nesting as in Python and other programming languages.

-- Each line begins with a keyword indicating a relationship type (Implication, Context,...) or ontological category (Visual, Experience,...) that is built into the NCE framework.

-- Terms such as $X and $Y denote variables, which are unbound by default but may become bound via nesting within (crisply, probabilistically, or fuzzily) quantified expressions.

-- Terms such as goto, box, etc. represent actions, entities or categories that the system knows about, either via initial programming, instruction or learning.

-- The relationship (Execution A B) denotes the execution of the procedure A on the argument or argument-list B.

-- The relationship (Evaluation A B) denotes the evaluation of the predicate A on the argument or argument-list B.

-- The relationship (Implication A B) denotes a probabilistic implication between expression A and expression B, which in the absence of quantification on the variables in A and B, has a truth value interpreted via averaging over all values for these variables, according to a probability distribution inferred from the system's experience.

-- A PredictiveImplication relationship denotes an Implication in which the first argument is constrained to occur before the second, in the calculation of the probabilistic truth value.

-- The SequentialAND relationship indicates a temporal sequencing of its arguments (which must be Execution relationships); if precise timing needs to be specified then more complex constructs must be used, but those won't be needed here.

-- The Similarity relationship connotes a probabilistic similarity between its arguments.

-- The Context relationship restricts all terms and relationships within the scope of its second argument, to the context defined by its first argument.

As all these relationship types may be graphically depicted as link-labels and all terms may be graphically depicted as node-labels, a nested relationship-set like the one depicted above may be graphically depicted as a network of labeled nodes and links, including links pointing to links. In this sense the knowledge representation used in the NCE, and illustrated in these examples, may be viewed as a rich kind of semantic net, which is mathematically a sort of weighted, labeled hypergraph.

Next, proceeding from icons to indices: Suppose one wanted to teach an agent that in order to get the teacher to give it a certain type of object, it should go to a box full of pictures and select a picture of an object that has commonly been used together with objects of that type, and bring it to the teacher. This is a combination of iconic and indexical semiosis, and would be achieved via the agent learning a rule of the form

```
Implication
  AND
    Context
      Visual
```

[1] A technical note on VAB knowledge representation and learning. The prior version of the VAB used with the AGISim virtual world used PLN as its primary learning mechanism; the current version used with Second Life uses MOSES as its primary learning mechanism. PLN-learned procedures are represented in the NCE AtomTable in a node-and-link format directly resembling the examples given here. On the other hand, MOSES-learned procedures are represented in the NCE as programs in an internal programming language called Combo. PLN-learned procedures are then translated into Combo for execution purposes; whereas MOSES-learned procedures are translated into logical Atoms for inference purposes. Here we depict all learned procedures using PLN-style, Atom notation, for sake of simplicity and consistency.

```
      Similarity $X $Z
    Context
      Experience
      SpatioTemporalAssociation $Z $Y
  PredictiveImplication
    SequentialAnd
      Execution goto box
      Execution grab $X
      Execution goto teacher
    Evaluation give me teacher $Y
```

Symbolism, finally, may be seen to emerge as a fairly straightforward extension of indexing. After all, how does an agent come to learn that a certain symbol refers to a certain entity? An advanced linguistic agent can learn this via explicit verbal instruction, e.g. one may tell it "The word 'hideous' means 'very ugly'." But in the early stages of language learning, this sort of instructional device is not available, and so the way an agent learns that a word is associated with an object or an action is through spatiotemporal association. For instance, suppose the teacher wants to teach the agent to dance every time the teacher says the word "dance" – a very simple example of symbolism. Assuming the agent already knows how to dance, this merely requires the agent learn the implication

```
PredictiveImplication
  SequentialAND
    Evaluation say teacher me "dance"
    Execution dance
  give teacher me Reward
```

And, once this has been learned, then simultaneously the relationship

```
SpatioTemporalAssociation dance "dance"
```

will be learned. What's interesting is what happens after a number of associations of this nature have been learned. Then, the system may infer a general rule of the form

```
Implication
  AND
    SpatioTemporalAssociation $X $Z
    HasType $X GroundedSchema
  PredictiveImplication
    SequentialAND
      Evaluation say teacher me $Z
      Execution $X
    Evaluation give teacher me Reward
```

This implication represents the general rule that if the teacher says a word corresponding to an action the agent knows how to do, and the agent does it, then the agent may get a reward from the teacher. Abstracting this from a number of pertinent examples is a relatively straightforward feat of probabilistic inference for the PLN inference engine.

Of course, the above implication is overly simplistic, and would lead an agent to stupidly start walking every time its teacher used the word "walk" in conversation and the agent overheard it. To be useful in a realistic social context, the implication must be made more complex so as to include some of the pragmatic surround in which the teacher utters the word or phrase $Z.

At this point we have experimented with making our AI system learn simple word-object associations and word-action associations, but haven't attempted to learn properly contextualized associations as would be useful in realistic social contexts – but we are optimistic that learning more fully contextualized knowledge of this sort is within the scope of our system, and look forward to experimenting with this during the coming year.

V. Teaching Gestural Communication

Based on the ideas described above, it is relatively straightforward to teach virtually embodied agents the elements of gestural comunication. This is important for two reasons: gestural communication is extremely useful unto itself, as one sees from its role in communication among young children and primates [22]; and, gestural communication forms a foundation for verbal communication, during the typical course of human language learning [23]. Note for instance the study described in [22], which "reports empirical longitudinal data on the early stages of language development," concluding that

> ...the output systems of speech and gesture may draw on underlying brain mechanisms common to both language and motor functions. We analyze the spontaneous interaction with their parents of three typically-developing children (2 M, 1 F) videotaped monthly at home between 10 and 23 months of age. Data analyses focused on the production of actions, representational and deictic gestures and words, and gesture-word combinations. Results indicate that there is a continuity between the production of the first action schemes, the first gestures and the first words produced by children. The relationship between gestures and words changes over time. The onset of two-word speech was preceded by the emergence of gesture-word combinations.

If young children learn language as a continuous outgrowth of gestural communication, perhaps the same approach may be effective for (virtually or physically) embodied AI's.

An example of an iconic gesture occurs when one smiles explicitly to illustrate to some other agent that one is happy. Smiling is a natural expression of happiness, but of course one doesn't always smile when one's happy. The reason that explicit smiling is iconic is that the explicit smile actually resembles the unintentional smile, which is what it "stands for."

This kind of iconic gesture may emerge in a socially-embedded learning agent through a very simple logic. Suppose that when the agent is happy, it benefits from its nearby friends being happy as well, so that they may then do happy things together. And suppose that the agent has noticed that when it smiles, this has a statistical tendency to make its friends happy. Then, when it is happy and near its friends, it will have a good reason to smile. So through very simple probabilistic reasoning, the use of explicit smiling as a communicative tool may result.

But what if the agent is not actually happy, but still wants some other agent to be happy? Using the reasoning from the prior paragraph, it will likely figure out to smile to make the other agent happy – even though it isn't actually happy.

Another simple example of an iconic gesture would be moving one's hands towards one's mouth, mimicking the movements of feeding oneself, when one wants to eat. Many analogous iconic gestures exist, such as doing a small solo part of a two-person dance to indicate that one wants to do the whole dance together with another person. The general rule an agent needs to learn in order to generate iconic gestures of this nature that, in the context of shared activity, mimicking part of a process will sometimes serve the function of evoking that whole process.

This sort of iconic gesture may be learned in essentially the same way as an indexical gesture such as a dog repeatedly drawing the owner's attention to the owner's backpack, when the dog wants to go outside. The dog doesn't actually care about going outside with the backpack – he would just as soon go outside without it – but he knows the backpack is correlated with going outside, which is his actual interest.

The general rule here is

```
R :=
Implication
  SimultaneousImplication
    Execution $X
    $Y
  PredictiveImplication
    $X
    $Y
```

I.e., if doing $X often correlates with $Y, then maybe doing $X will bring about $Y. This sort of rule can bring about a lot of silly "superstitious" behavior but also can be particularly effective in social contexts, meaning in formal terms that

```
Context
  near_teacher
  R
```

holds with a higher truth value than R itself. This is a very small conglomeration of semantic nodes and links yet it encapsulates a very important communicational pattern: that if you want something to happen, and act out part of it – or something historically associated with it -- around your teacher, then the thing may happen.

Many other cases of iconic gesture are more complex and mix iconic with symbolic aspects. For instance, one waves one hand away from oneself, to try to get someone else to go away. The hand is moving, roughly speaking, in the direction one wants the other to move in. However, understanding the meaning of this gesture requires a bit of savvy or experience. One one does grasp it, however, then one can understand its nuances: For instance, if I wave my hand in an arc leading from your direction toward the direction of the door, maybe that means I want you to go out the door.

Purely symbolic (or nearly so) gestures include the thumbs-up symbol mentioned above, and many others including valence-indicating symbols like a nodded head for YES, a shaken-side-to-side head for NO, and shrugged shoulders for "I don't know." Each of these valence-indicating symbols actually indicates a fairly complex concept, which is learned from experience partly via attention to the symbol itself. So, an agent may learn that the nodded head corresponds with situations where the teacher gives it a reward, and also with situations where the agent makes a request and the teacher complies. The cluster of situations corresponding to the nodded-head then forms the agent's initial concept of "positive valence," which encompasses, loosely speaking, both the good and the true.

Summarizing our discussion of gestural communication: An awful lot of language exists between intelligent agents even if no word is ever spoken. And, our belief is that these sorts of non-verbal semiosis form the best possible context for the learning of verbal language, and that to attack verbal language learning outside this sort of context is to make an intrinsically-difficult problem even harder than it has to be. And this leads us to the final part of the paper, which is a bit more speculative and adventuresome. The material in this section and the prior ones describes experiments of the sort we are currently carrying out with our virtual agent control software. We have not yet demonstrated all the forms of semiosis and non-linguistic communication described in the last section using our virtual agent control system, but we have demonstrated some of them and are actively working on extending our system's capabilities. In the following section, we venture a bit further into the realm of hypothesis and describe some functionalities that are beyond the scope of our current virtual agent control software, but that we hope to put into place gradually during the next 1-2 years. The basic goal of this work is to move from non-verbal to verbal communication.

VI. Teaching verbal communication

A purist approach to endowing embodied virtual agents with linguistic facility would be to provide them with zero "linguistic hard-wiring" and have them learn to communicate linguistically entirely based on (embodied, interactive) experience. Whether human language learning is pure in this sense is a matter of some contention, as human psycho- and neuro-linguistics are not yet advanced enough to tell us what kind (if any) of linguistic knowledge comes wired into the human brain [24]. But in any case, our current suspicion is that this would not be the optimal path to creating intelligent virtual agents with robust linguistic facility. Rather, it seems more practical to create virtual agents with in-built linguistic capability that is designed and programmed with ongoing experience-based adaptation in mind. This approach requires significant care, but also has dramatic potential for acceleration of progress, because of the wide variety of powerful computational linguistics tools that have been developed over the last few years.

A simplistic dichotomous analysis of computational linguistics frameworks would divide them into statistical learning approaches versus expert-rule-based approaches. So far, pure statistics based approaches have proved effective for information retrieval applications and for some specific linguistic tasks such as word sense disambiguation. Furthermore, some impressive systems have been built in which purely statistical or machine learning algorithms are used to determine appropriate conversational responses to linguistic inputs [5,25].

However, if one wishes to create a system that carries out abstract reasoning based on information gained from language, currently the only workable approach is to use an expert-rule-based approach, involving a syntax parser with hand-built rules, that maps sentences into logical relationships (and inversely, though less work has been done on this, a language generator with hand-built rules that maps logical relationships into sentences). Now, it may be argued that explicit logical reasoning is the wrong approach to AI altogether, and that one should instead build out and train a neural net or some other sort of purely subsymbolic adaptive learning system to the point where it can implicitly carry out functions similar to those we call logical reasoning. But realistically, the current state of the art seems quite far from this.

Our approach in the NCE has been to combine symbolic and subsymbolic representations and learning methods; and this means that in the linguistic domain, though we are able to make use of statistical and machine learning methods, we are also eager to make use of expert-rule-based methods with their ability to speak directly to the formal-logic-based aspect of the NCE. As it happens, the judicious combination of statistical and expert-rule-based methods is increasingly common in the computational linguistics world, so our deployment of this sort of combination in the NCE context is not all that unusual. As an example of the rules+statistics combination in the computational linguistics literature, Dekang Lin's MiniPar system applies expert rules to carry out parsing, but then uses statistical methods to rank the parses produced by the rule-based parser, and uses statistical analysis of the parses produced by the parser to automatically divide words into semantic categories [26,27].

Of all the computational linguistics approaches in the literature, Word Grammar [28] is probably the closest one to the NCE in overall philosophy, and the most amenable to experiential adaptive learning. However, Word Grammar is not currently very mature as a computational framework, and so we are making use of other tools, though deploying them in a somewhat Word Grammar like way.

Our current computational linguistics framework contains the following components, some original and some created by others and integrated into our framework:

-- 1. An *entity extractor* drawn from the GATE [29] framework

-- 2. The *link parser*, a dependency grammar parser [30], which relies on an extensive association of words with syntactic link types called the "link grammar dictionary"

-- 3. *RelEx*, a rule-based system mapping the output of the link parser into higher-level semantic relationships

-- 4. *RelEx2Frame*, a rule-based system mapping the output of RelEx into frame-element relationships (some directly drawn from FrameNet [31], some created analogously to those in FrameNet)

-- 5. A statistical *parse ranker*, that ranks the parses output by RelEx based on probabilities estimated for their component links according to frequency in a corpus

-- 6. Word *sense disambiguation and reference resolution* algorithms based on statistical analysis of the RelEx interpretations of the sentences in a corpus

-- 7. *RelExpres*s, a language generation system that uses the link grammar dictionary to translate RelEx output into natural language sentences. Components 1-6 mentioned above have to do with language comprehension, but many have correlates within RelEx press: e.g. word-selection and reference-insertion are the correlates of word sense disambiguation and reference resolution; expression ranking is the correlate of parse ranking. The algorithms used for each comprehension component and its generation correlate are related and there is often significant shared software.

Components 1-3 are fairly mature and have been used by Novamente LLC in commercial projects; see [32] for a description of a prototype application using RelEx, the link parser, PLN and specialized entity extractors to carry out simple inferences on knowledge extracted from PubMed research abstracts. Components 4-5 are less mature, and components 5-7 are still in an early stage of software development, and not fully functional. In order to be used for conversation, these components must be controlled by a "discourse management" subsystem, which in the NCE is integrated with the overall NCE action-selection framework; this subsystem includes various procedures that deal with linguistic-pragmatics issues such as conversational implicature, and also falls into the "early stage" category.

It is interesting to enumerate the aspects in which each of the above components appears to be capable of tractable adaptation via experiential, embodied learning:

-- 1. Words and phrases that are found to be systematically associated with particular objects in the world, may be added to the "gazeteer list" used by the entity extractor

-- 2. The link parser dictionary may be automatically extended. In cases where the agent hears a sentence that is supposed to describe a certain situation, and realizes that in order for the sentence to be mapped into a set of logical relationships accurately describing the situation, it would be necessary for a certain word to have a certain syntactic link that it doesn't have, then the link parser dictionary may be modified to add the link to the word. (On the other hand, creating new link parser link *types* seems like a very difficult sort of learning – not to say it is unaddressable, but it will not be our focus in the near term.)

-- 3. Similar to with the link parser dictionary, if it is apparent that to interpret an utterance in accordance with reality a RelEx rule must be added or modified, this may be automatically done. The RelEx rules are expressed in the format of relatively simple logical implications between Boolean combinations of syntactic and semantic relationships, so that learning and modifying them is within the scope of a probabilistic logic system such as Novamente's PLN inference engine.

-- 4. The rules used by RelEx2Frame may be experientially modified quite analogously to those used by RelEx

-- 5. Our current statistical parse ranker ranks an interpretation of a sentence based on the frequency of occurrence of its component links across a parsed corpus. A deeper approach, however, would be to rank an interpretation based on its commonsensical plausibility, as inferred from experienced-world-knowledge as well as corpus-derived knowledge. Again, this is within the scope of what an inference engine such as PLN should be able to do.

-- 6. Our word sense disambiguation and reference resolution algorithms involve probabilistic estimations that could be extended to refer to the experienced world as well as to a parsed corpus. For example, in assessing which sense of the noun "run" is intended in a certain context, the system could check whether stockings, or sports-events or series-of-events, are more prominent in the currently-observed situation. In assessing the sentence "The children kicked the dogs, and then they laughed," the system could map "they" into "children" via experientially-acquired knowledge that children laugh much more often than dogs.

-- 7. RelExpress uses the link parser dictionary, treated above, and also uses rules analogous to (but inverse to) RelEx rules, mapping semantic relations into brief word-sequences. The "gold standard" for RelExpress is whether, when it produces a sentence S from a set R of semantic relationships, the feeding of S into the language comprehension subsystem produces R (or a close approximation) as output. Thus, as the semantic mapping rules in RelEx and RelEx2Frame adapt to experience, the rules used in RelExpress must adapt accordingly, which poses an inference problem unto itself.

All in all, when one delves in detail into the components that make up our hybrid statistical/rule-based NLP system, one sees there is a strong opportunity for experiential adaptive learning to substantially modify nearly every aspect of the NLP system, while leaving the basic framework intact.

This approach, we suggest, may provide means of dealing with a number of problems that have systematically vexed existing linguistic approaches. One example is parse ranking for complex sentences: this seems almost entirely a matter of the ability to assess the semantic plausibility of different parses, and doing this based on statistical corpus analysis seems unreasonable. One needs knowledge about a world to ground reasoning about plausiblity.

Another example is preposition disambiguation, a topic that is barely dealt with at all in the computational linguistics literature (see e.g. [33] for an indication of the state of the art). Consider the problem of assessing which meaning of "with" is intended in sentences like "I ate dinner with a fork", "I ate dinner with my sister", "I ate dinner with dessert." In performing this sort of judgment, an embodied system may use knowledge about which interpretations have matched observed reality in the case of similar utterances it has processed in the past, and for which it has directly seen the situations referred to by the utterances. If it has seen in the past, through direct embodied experience, that when someone said "I ate cereal with a spoon," they meant that the spoon was their tool not part of their food or their eating-partner; then when it hears "I ate dinner with a fork," it may match "cereal" to "dinner" and "spoon" to "fork" (based on probabilistic similarity measurement) and infer that the interpretation of "with" in the latter sentence should also be to denote a tool.

How does this approach to computational language understanding tie in with gestural and general semiotic learning as we discussed earlier? The study of child language has shown that early language use is not purely verbal by any means, but is in fact a complex combination of verbal and gestural communication [23]. With the exception of point 1 (entity extraction) above, every one of our instances of experiential modification of our language framework listed above involves the use of an understanding of what situation actually exists in the world, to help the system identify what the logical relationships output by the NLP system are supposed to be in a certain context. But a large amount of early-stage linguistic communication is social in nature, and a large amount of the remainder has to do with the body's relationship to physical objects. And, in understanding "what actually exists in the world" regarding social and physical relationships, a full understanding of gestural communication is important. So, the overall pathway we propose for achieving robust, ultimately human-level NLP functionality is as follows:

-- 1. The capability for learning diverse instances of semiosis is established

-- 2. Gestural communication is mastered, via nonverbal imitative/reinforcement/corrective learning mechanisms such as we are now utilizing for our embodied virtual agents

-- 3. Gestural communication, combined with observation of and action in the world and verbal interaction with teachers, allows the system to adapt numerous aspects of its initial NLP engine to allow it to more effectively interpret simple sentences pertaining to social and physical relationships

-- 4. Finally, given the ability to effectively interpret and produce these simple and practical sentences, probabilistic logical inference allows the system to gradually extend this ability to more and more complex and abstract senses, incrementally adapting aspects of the NLP engine as its scope broadens.

Our current work focuses on steps 1 and 2, and on building out and tuning the initial NLP engine to be used in step 3. While not a "pure experiential learning" methodology, we believe this is a pragmatic approach which stands a reasonably high chance of success in the relatively near term – utilizing the power of experiential learning, but rather than making it learn everything from scratch, allowing it to use appropriately architected computational-linguistics tools as an initial condition.

VII. THE ROLE OF NON-VERBAL UTTERANCES AND PHONOLOGY IN LANGUAGE LEARNING

In this brief section we will mention another potentially important factor that we have intentionally omitted in the above analysis – but that may wind up being very important, and that can certainly be taken into account in our framework if this proves necessary. We have argued that gesture is an important predecessor to language in human children, and that incorporating it in AI language learning may be valuable. But there is another aspect of early language use that plays a similar role to gesture, which we have left out in the above discussion: this is the acoustic aspects of speech.

Clearly, pre-linguistic children make ample use of communicative sounds of various sorts. These sounds may be iconic, indexical or symbolic; and they may have a great deal of subtlety. Steven Mithen [34] has argued that non-verbal utterances constitute a kind of proto-language, and that both music and language evolved out of this. Their role in language learning is well-documented also [35].

We are uncertain as to whether an exclusive focus on text rather than speech would critically impair the language learning process of an AI system. We are fairly strongly convinced of the importance of gesture because it seems bound up with the importance of semiosis – gesture, it seems, is how young children learn flexible semiotic communication skills, and then these skills are gradually ported from the gestural to the verbal domain. Semiotically, on the other hand, phonology doesn't seem to give anything special beyond what gesture gives. What it does give is an added subtlety of emotional expressiveness – something that is largely missing from virtual agents as implemented today, due to the lack of really fine-grained facial expressions. Also, it provides valuable clues to parsing, in that groups of words that are syntactically bound together are often phrased together acoustically.

If one wished to incorporate acoustics into the framework described above, it would not be objectionably difficult on a technical level. Speech-to-text [36] and text-to-speech software [37] both exist, but neither have been developed with a view specifically toward conveyance of emotional information. One could approach the problem of assessing the emotional state of an utterance based on its sound as a supervised categorization problem, to be solved via supplying a machine learning algorithm with training data consisting of human-created pairs of the form (utterance, emotional valence). Similarly, one could tune the dependence of text-to-speech software for appropriate emotional expressiveness based on the same training corpus. This would represent significant additional effort but would not be as difficult as some other aspects of the programme described here.

In sum, this is a direction that we are open to explore, yet uncertain of its necessity, so depending on various practical considerations we may wind up deferring it until/unless its critical nature becomes clearer via theory or experimentation.

CONCLUSION

Beginning with our current applied work creating virtual animals that learn non-verbal behaviors via interacting with human-controlled agents in Second Life and other virtual worlds, we have traced a path via which virtual embodiment may potentially be used to enable AI systems to acquire robust linguistic communication faculties. There are many steps along this path, and hence many potential points of failure; but we believe this is the most cognitively and computationally plausible pathway yet articulated for transitioning from the current primitive state of AI systems into the desired future state wherein it is possible to communicate complex ideas with AI systems using natural language. As we view robust NL communication as the Rubicon which, once crossed, will allow the pace of AI development to accelerate extremely rapidly, we view this as an extremely important train of thought and endeavor.

REFERENCES

[1] R. Kurzweil. The Singularity Is Near. Penguin Press

[2] Goertzel, Ben and Cassio Pennachin, Editors. Artificial General Intelligence. Springer, New York

[3] See http://www.engagingexperience.com/2006/07/ai50_first_poll.html for a poll taken of attendees at the AI@50 conference

[4] Samsonovich, A. V. (2006). Biologically inspired cognitive architecture for socially competent agents. In Upal, M. A., & Sun, R. (Eds.). Cognitive Modeling and Agent-Based Social Simulation: Papers from the AAAI Workshop. AAAI Technical Report, vol. WS-06-02, pp. 36-48. Menlo Park, CA: AAAI Press

[5] R. Hecht-Nielsen R (2005) Cogent confabulation. Neural Networks 18:111-115

[6] W Duch, Oentaryo R.J, Pasquier M, Cognitive architectures: where do we go from here?, Proceedings of AGI 2008, IOS Press

[7] P Wang, Rigid Flexibility: The Logic of Intelligence, Sppringer

[8] Shapiro, S. An Introduction to SNePS 3. In Bernhard Ganter & Guy W. Mineau, Eds. Conceptual Structures: Logical, Linguistic, and Computational Issues. Lecture Notes in Artificial Intelligence 1867. Springer-Verlag, Berlin, 2000, 510-524.

[9] B Goertzel and Cassio Pennachin (2004). Novamente: An Integrative Approach to Artificial General Intelligence, in Goertzel, Ben and Cassio Pennachin, Editors. Artificial General Intelligence. Springer, New York

[10] B Goertzel, Moshe Looks and Cassio Pennachin. Novamente: An Integrative Architecture for Artificial General Intelligence. Proceedings of AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, Washington DC, August 2004

[11] B Goertzel. Virtual Easter Egg Hunting. In Goertzel, Ben and Pei Wang, Eds. (2007). Advances in Artificial General Intelligence. IOS Press.

[12] M. Ikle' and Ben Goertzel (2008). Probabilistic Quantifier Logic for General Intelligence: An Indefinite Probabilities Approach, Proceedings of AGI 2008, IOS Press

[13] M. Ikle', Ben Goertzel, Izabela Goertzel and Ari Heljakka (2007). Indefinite Probabilities for General Intelligence, in Advances in Artificial General Intelligence, IOS Press.

[14] M Looks. Competent Program Evolution. PhD Thesis, Department of Computer Science, Washington University, St. Louis, 2006

[15] M. Looks. Indefinite Probabilities for General Intelligence, in Advances in Artificial General Intelligence, IOS Press, 2007

[16] B. Goertzel, Cassio Pennachin, Nil Geissweiller, Moshe Looks, Andre Senna, Welter Silva, Ari Heljakka, Carlos LopesAn Integrative Methodology for Teaching Embodied Non-Linguistic Agents, Applied to Virtual Animals in Second Life

[17] B Goertzel. Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity, AI Journal, 2008, to appear

[18] A Turing. "Computing machinery and intelligence", Mind LIX (236): 433-460, 1950

[19] Stan Franklin, Uma Ramamurthy, Sidney K. D'Mello, Lee McCauley, Aregahegn Negatu, Rodrigo Silva L., and Vivek Datla. 2007 to appear. LIDA: A computational model of global workspace theory and developmental learning. In AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches. Arlington, VA: AAAI.

[20] A Heljakka, Ben Goertzel, Welter Silva, Izabela Goertzel and Cassio Pennachin. Reinforcement Learning of Simple Behaviors in a Simulation World Using Probabilistic Logic, in Advances in Artificial General Intelligence, IOS Press. 2007

[21] C. Peirce. Collected papers: Volume V. Pragmatism and pragmaticism. Cambridge, MA, USA: Harvard University Press, 1934

[22] O Capirci, Annarita Contaldo, Maria Cristina Caselli and Virginia Volterra . From action to language through gesture: A longitudinal perspective, Gesture 5:2, 155-157 (2005)

[23] K. Emmory and J. Reilly. Language, Gesture and Space. Lawrence Erlbaum, 1995

[24] L Jenkins, Editor. Variation and Universals in Biolinguistics. Elsevier Science, 2004

[25] A Borzenko. A Cognitive Substrate for Natural Language Understanding, Proceedings of AGI 2008, IOS Press

[26] D. Lin, 1998. Dependency-based Evaluation of MINIPAR. In Workshop on the Evaluation of Parsing Systems, Granada, Spain, May, 1998.

[27] D. Lin and P. Pantel. 2001. Induction of Semantic Classes from Natural Language Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001. pp. 317-322

[28] R Hudson. Language Networks. The new Word Grammar. Oxford University Press

[29] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[30] D Sleator and D Temperley. Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

[31] C F Baker., Fillmore, Charles J., and Lowe, John B. (1998): The Berkeley FrameNet project. In Proceedings of the COLING-ACL, Montreal, Canada

[32] B Goertzel, Hugo Pinto, Cassio Pennachin et al. Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts. Proceedings of the BioNLP Workshop/HLT-NAACL 2006

[33] Kenneth C. Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, pages 24-29

[34] S Mithen. The Singing Neanderthals, Harvard University Press, 2007

[35] J. Morgan, Katherine Demuth. Bootstrapping From Speech To Grammar in Early Acquisition. Lawrence Erlbaum, 1995

[36] X Huang. Spoken Language Processing. Prentice-Hall 2001

[37] M Tatham and K Morton. Developments in Speech Synthesis. Wiley, 2005.