Towards "Natural" Natural Language in Machine Cognition

Pentti O. A. Haikonen

pentti.haikonen@nokia.com

Abstract

Machine understanding of language would obviously offer numerous business opportunities such as truly conversational human-robot interfaces, conversational information systems, automated text understanding and summarization, information search, junk mail filtering and eventually true thinking machines. Traditionally the problem of automated language understanding has been approached from the linguistic point of view; the utilization of formal syntax and lexicons. So far, the results have left much to be desired. Obviously the brain does not process language in this way and there must be another, natural way to do it.

It is proposed that machine understanding of a natural language should be based on the emulation of the cognitive processes of the brain instead of preprogrammed formal syntax and vocabulary. Meaning should be grounded to real world entities. Same processes for perceiving and understanding the world and linguistic understanding should be used, no specific linguistic machinery is assumed to exist. In the brain sensory processes provide information about the world and this information is integrated into a continuously updated multimodal "situation model". A natural language is seen as means for the description and evocation of these inner "situation models". The author's proposition for the artificial implementation of these principles is called "The multimodal model of language" and its main application area would be the use and understanding of natural language in cognitive machines and robots.

Keywords: Machine cognition, natural language, situation models, cognitive architectures, grounding of meaning, understanding.

Introduction

We all know, understand and can use at least one natural language. Therefore we should also be able to tell how a language works and consequently, machine implementation of language use and understanding should be easy. However, experience shows that this is not the case.

What takes place in the brain and in the mind when a story is being heard or read and understood? These processes are not directly available for introspection and we must find these in indirect ways. For instance, hearing, reading and understanding a story does not involve the learning of the story as strings of words by heart, yet something must be memorized and remembered. We do not learn stories by heart without extra effort, but nevertheless we will learn and remember the gist of the story with only one reading. Thereafter we are also able to paraphrase the story. Thus, a reader must extract the information contained in the text, but how does this happen?

The understanding of text and stories can be demonstrated by asking questions about the subject matter of the text. If these questions cannot be answered, then obviously no understanding has taken place. It should also be noted that full-scale understanding would involve extensive background knowledge, which would supply a contextual framework and any necessary information that is not included in the story itself.

At first sight, the obvious way of realizing text understanding by a computer would involve the storing of the text in the computer memory as a data file. Then the stored text could be scanned and the requested information could be extracted by pattern matching, statistical or other algorithms. This approach, when properly executed, will lead to initial success. However, at the end this approach turns out to be naïve and not scaling well for arbitrary texts. This is definitely not the way that is utilized by the brain, either. It turns out that the treatment of language as a separate autonomous system will not suffice. The meaning of words and sentences cannot be defined ultimately by other words and sentences as this would only lead to circular definitions. Therefore references to real world would be necessary; grounding of meaning must be incorporated.

The brain perceives the world by a variety of sensors and these percepts are processed in ways that facilitate meaningful action in the world. Therefore the observed world manifests itself to us as situations; objects, entities and actions at different locations. These situations have meaning, we know what the objects, entities and actions are and what we could or should do about them; we understand the situation. The perceived situations are subjectively interpreted impressions of the world, kinds of continuously updated “situation models”. These “situation models” or “scenes” are mental and therefore can also be induced in lesser vividness internally, by imagination.

Zwaan and Radvansky (1998) argue that situation models are necessary for language comprehension and they lament that traditionally cognitive psychologists used to see text comprehension as the mental construction and retrieval of the text itself rather than the described situation. The author agrees with Zwaan and Radvansky and sees a natural language as means for the description and evocation of these inner “situation models” or “scenes”. This view does not exclude a certain degree of autonomy of language and hence the successful execution of limited “language understanding” without any actual grounding of meaning. However, this view necessarily calls for the utilization of “situation models” in full-blown artificial natural language understanding systems.

There have been some attempts towards the use of grounded language in cognitive machines and robots. The author has devised a simple neural system with real video camera that was able to use and understand simple natural language sentences and learn to recognize and answer questions based on internally evoked imagery of objects (Haikonen 1999). Mavridis and Roy (2005) have done something similar but by traditional AI means. They describe a grounded situation model for robots, which bridges language, perception and action in a way that allows the robot to answer questions about situations and imagine verbally described situations.

Mental Situation Models

We can focus our sharpest visual attention to one object at a time, but at the same time we are retaining the information of the locations and motions of the other objects, also those ones that are behind us. In this way we have a kind of “map”, “scene” or “situation model” of our environment. However, this situation model is not limited to the immediate spatial aspects of the surroundings as it involves and engages also, for instance, a short piece of history, associative linking to long-term memory, background information and emotional evaluation. It has also some predictive power based on this additional information.

In order to create this situation model the brain has to associate the objects, properties, locations and actions with each other. This situation model is necessarily multimodal, containing pertinent information from various sensors as well as from memory. The brain can sustain this situation model, reset and update it when necessary and retrieve information from it as necessary. This situation model changes when the situation changes; therefore it is a kind of running model and could also be called "running world model" in the style of Sommerhof (2000). Past situation models (memories) are also accessible. However, these are not supported by present percepts and are therefore understood to represent past events (or imagined ones).

Language is linked to situation models. Verbal descriptions, sentences and stories, evoke situation models, which are then processed normally. These verbally evoked situation models do not have to be as vivid as those evoked by perception processes. Minimal relational information is enough for the paraphrasing of the situation. The understanding of the situation calls for the inspection of the model, "the scene". Verbal questions direct attention on certain details of the model. Therefore it can be seen that words have also the function of focusing attention (this idea is expressed also in Marchetti 2006).

The neural machinery of the brain is adapted to execute the functions that are required to support this process. The author proposes that the same functions and neural architecture are used and are sufficient for the utilization of language. However, this is not to say that more neurons were not necessary; instead, additional neural capacity would be required, but this would only repeat the general architecture.

Implications to Artificial Cognitive Architectures

The human cognitive system consists of the brain and the sensing system. The sensing system may be divided into separate modalities, such as visual, auditory, touch etc. modalities. Important sensory modalities are also those that sense the position of the various body parts, such as hands, fingers and so on. This information is relevant in connection with motor commands as they allow the feedback control of body part motion. Each sensory modality must then consist of a module that has input stages that receive the signals from the sensor and output stages that generate some response. The modules must communicate with each other. For example, the visual module must transmit its information to hand position detection module so that the hand could be commanded to reach out for a visually detected object or, likewise, to draw a picture of the visually detected object. This leads to another requirement for the modules; they must have intermediate stages that are able to form associative connections with other modules. Yet another architectural feature is necessary, namely feedback. This feedback shall translate the internally generated response into a corresponding percept of the same sensory modality. This feedback would then allow the introspection of mental content in the terms of sensory percepts. This operation would also facilitate the act of drawing from memory; plotting mental imagery from the visual module as if it were a currently perceived external visual object.

The basic requirement for language in a cognitive system is seen as the following. The flow of sensory percepts depicting a given situation must be able to evoke a corresponding sequence of linguistic symbols; words that name entities, show their relationships as well as action and necessary syntactic devices. Also, the linguistic expressions must be able to evoke inner activity that is approximately equivalent to the activity that would result from the actual perception of the described situation. If we realize that the "linguistic symbols" and "sentences" are actually sequences of auditory percepts then we should see that the basic requirement for language calls for associative connections between, at least, the visual and auditory modules; a requirement that is already satisfied by the more general requirements of cognitive architectures. It should be obvious that we are able to name and describe percepts and events that are sensed by sensory modalities other than the visual one, too. This would be facilitated by the associative cross-connections between the auditory module and the other sensory modules.

The Multimodal Model of Language

The multimodal model of language (Haikonen 2003, 2007) is devised for the neural implementation of natural language in artificial cognitive systems, especially those that have the faculty of multisensory perception.

The implementation of the multimodal model of language calls for a neuron group assembly, a plane, for each sensory modality. Examples of these are the visual, auditory, touch etc. sensory modalities. These neuron group assemblies store and associatively manipulate representations of their own kind. For instance, the visual plane can handle representations of visual objects, their relative locations, motion, change, etc. The auditory plane can handle representations of sound patterns, their pitch, rhythm, duration, change, etc. The touch plane can handle representations from the cutaneous receptors. The representations within a neuron group assembly can be associated with each other. Associative cross-connections between the sensory modality planes are also possible. In this way, for instance, a visual percept may evoke a sound percept and vice versa.

The important point here is the possibility to set certain sound patterns, “words” to represent arbitrary entities that they do not naturally represent. This is achieved via modified associative Hebbian learning. In this way certain neural activation patterns become used as symbols with meanings that go beyond their original causal meanings, namely the original auditory stimuli. The same mechanism allows also the use of written language.

The proper implementation of the multimodal model of language involves the grounding of meaning of words and sentences to external world entities and actions and the use of situation models. Thus the model will also allow the verbal description of external and imagined situations and the evocation of inner situation models for the verbally described situations.

References

Haikonen, P. O. (1999). *An Artificial Cognitive Neural System Based on a Novel Neuron Structure and a Reentrant Modular Architecture with Implications to Machine Consciousness*. Dissertation for the degree of Doctor of Technology, Helsinki University of Technology, Applied Electronics Laboratory, Series B: Research Reports B4

Haikonen, P. O. (2003). *The Cognitive Approach to Conscious Machines*. UK: Imprint Academic.

Haikonen, P. O. (2007). *Robot Brains; Circuits and Systems for Conscious Machines*. UK: John Wiley & Sons, Ltd.

Marchetti, G. (2006). A presentation of Attentional Semantics. *Cognitive Processing* Vol. 7, No. 3, 2006

Mavridis, N., Roy, D. (2005) *Grounded Situation Models for Robots: Bridging Language, Perception, and Action*. Retrieved on 14.4.2008 from www.media.mit.edu/cogmac/publications/WS1105MavridisN.pdf

Sommerhof, G. (2000). *Understanding Consciousness*. London: Sage Publications.

Zwaan, R. A., Radvansky, G. A. (1998) Situation Models in Language Comprehension and Memory. *Psychological Bulletin* Vol. 123, No. 2, 1998, 162-185