MEANING-CENTRIC FRAMEWORK FOR NATURAL TEXT/SCENE UNDERSTANDING BY ROBOTS

M. XIE∗, J. S. KANDHASAMY and H. F. CHIA

School of Mechanical & Production Engineering
Nanyang Technological University
Nanyang Avenue, 639798, Singapore
∗mmxie@ntu.edu.sg

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

In the past fifty years, efforts in classical AI have focussed on computerizing human intelligence. Naturally, computerized human intelligence is not a proof of machine or robot intelligence because the programs underlying computerized human intelligence are still made by humans. So far, there is no computer nor robot which is creative enough to master its own language and to compose a text expressing its intentions. Thus, it is time to shift our research focus from computerizing human intelligence to developing machine intelligence. A first and necessary step towards this goal is to make machines or robots learn, manipulate, understand and create both elementary and composite meanings encoded in a natural language such as English. Elementary and composite meanings could be acquired through both sample texts and images. Hence, we propose a learning-synthesis-analysis framework which aims to enable a robot, or computer, to understand and convey meaning through texts or images. The main contribution of this paper is to lay out a sound foundation on which interdisciplinary research could effectively progress toward the development of machines or robots that understand meaning through texts or images.

Keywords: Robot intelligence; natural text understanding; natural scene understanding.

1. Introduction

It is our position and belief that robot intelligence is not equal to human intelligence. Most importantly, the discovery from the study of text and scene understanding by humans may not necessarily lead to engineering solutions to the problem of text and scene understanding by robots.

1.1. Essence of interdisciplinary research

Although it may appear obvious, it is worth repeating that the contemporary history of scientific research is characterized by the interplay between discovery-centric research and invention-centric research, as outlined in Fig. 1.
The results obtained from the discovery of how nature works is certainly the engine which propels the creation of universal knowledge. The accumulation of knowledge, in turn, drives our curiosity and desire to undertake the creation of new things. And, systematically, the imperfection of man-made inventions inspires us either to look for better solutions in making new things, or to find new inspiration from the nature. Obviously, the results of invention will benefit, and encourage, discovery-centric research for the sake of achieving a better understanding of the nature.

1.2. Emergence of humanoid robots

The advances of science and technology is making products and services increasingly sophisticated. As a result of the emergence of smart products in society and intelligent machines in industry, an obvious question arises: how should we harmonize the co-existence of humans and machines (or tools) in industry and society? In other words, is it desirable to make humans behave like machines because of the adherence to man-made rules, procedures and (programmable) instructions? Or, should we make machines behave like humans because of the embodiment of human-like psychology and intelligence? Undoubtedly, it makes sense to expect future machines to have not only a body for action execution, but also an artificial mind which could learn and engender human-like, or human-instructed, behaviors.\(^1,2\)
It has been an ambition of humans to develop human-like machines, called robots, in general, or humanoid robots, in particular. Over the past decades, tremendous progress has been made in engineering for the understanding of design, analysis and control of grasping, biped walking and manipulating mechanisms. 3–5 In parallel, a great many interesting results have been generated from the investigation of learning, and comprehension of language and images by human beings.8,18,22

Fuelled by the excitement generated by the impressive shows of HONDA’s and SONY’s humanoid robots, should we say that the great new era of humanoid robots has just started, and that more excitement awaits us?

Humanoid robots are not natural creatures. They are man-made inventions. Therefore, the development of humanoid robots is part of invention-centric research, which should not be overshadowed by our limited understanding of nature. In fact, we have full freedom to exercise our creativity in the invention of humanoid robots. Of course, humanoid robot research will benefit a lot from the results of the studies on how the human body and mind work. As the outcome of invention will stimulate discovery-centric research in various ways, the humanoid robot, which is an embodiment of mind and body, is undoubtedly an interesting platform to validate, or apply, theories as a result of studies in neuroscience, psychology, learning, and cognition.

The synergetic integration of mechanics, electronics, control, communication, perception, decision-making, artificial psychology, and artificial intelligence has greatly enlarged the scope of scientific investigation into the engineering approaches and principles underlying the development of humanoid robots. As both discovery-centric and invention-centric research in this field makes progress side-by-side, a new horizon is coming into view on which fruitful results are expected to emerge in the forms of various new theories and products. Compared to industrial robot research, humanoid robot research certainly offers greater opportunities and inspiration for new invention and discovery.

This paper is organized as follows: Sec. 2 discusses the issue of what intelligence is. A new definition with an engineering flavor is proposed in order to obtain a durable definition. Section 3 addresses the issue of a robot’s mental architecture. In this section, the concepts of the real and mental worlds are described. Section 4 proposes a meaning-centric framework for natural text understanding by robots. A software agent called KnowNet is briefly introduced with some preliminary results of our ongoing research. Section 5 briefly discusses a meaning-centric framework for natural scene understanding by robots. A detailed conclusion is given in Sec. 6.

2. Intelligence

Modern society is fuelled by a knowledge-based economics in which knowledge becomes a commodity that can be traded. There is no doubt that the creation of knowledge requires intelligence. Hence, we may ask the question: “What is intelligence?” Although the literature contains numerous attempts to find a
durable definition of intelligence, few consider the engineering aspects of the term “intelligence.” Without the consideration of the engineering aspects, which promote objectivity instead of subjectivity, one may equally face the difficulty of credibly defining the term “intelligence” as well as the term “motion” (i.e. what is motion?).

2.1. Human intelligence

What is human intelligence? Unfortunately, there is no credible consensus on the definition. Arising from common sense, various suggested definitions have emerged. Here are some examples which consider that human intelligence is not a function of time:

- “Intelligence is an innate quality, as distinct from abilities acquired through learning” (Herbert Spencer);
- “Intelligence is the innate general cognitive ability” (Cyril Burt);
- “Intelligence is an inborn or innate quality, as distinct from abilities acquired through individual experience” (from Encyclopedia Britannica).

Obviously, many people will disagree with these definitions because of the absence of the dimension of time. To remedy these deficiencies due to the lack of consideration of time (i.e. learning, or experience), some other definitions have been put forward. These include:

- “Intelligence is the ability to carry out abstract thinking” (L. M. Terman);
- “Intelligence is the capacity to act purposefully, think rationally and deal effectively with the environment” (D. Wechsler).

2.2. Computerized human intelligence

With the advent of computers, researchers and scientists are excited about the possibility of computerizing human intelligence in the form of computer programs so as to make computers possess a certain degree of human-like intelligence. As a result of this human endeavor, a new technical term was born, that is: Artificial Intelligence (AI).

Interestingly enough, again, researchers and scientists could not agree on a common definition of AI. In the literature, some definitions are:

- “AI is the study of the computations that make it possible to perceive, reason, and act” (Winston);
- “AI is the study of mental faculties through the use of computational models” (Charniak and McDermott);
- “AI is the study of how to make computers do things at which, at the moment, people do better” (Rich and Knight);
- “AI is the art of creating machines that perform functions that require intelligence when performed by people” (Kurzweil).

The study of AI historically focuses on problem-solving (i.e. analysis) and learning. The difficult issue of synthesis (i.e. creativity) has not yet received much
attention. Although artificial intelligence literally means man-made intelligence in a machine, the content discussed under artificial intelligence suggests that it actually deals with partly computerized human intelligence or computational rationality.

If AI is about computerizing human intelligence, its origin should go back to the early days when the computer was invented (the 1940s). However, the officially recognized date of birth of AI is 1956, which was marked by the Dartmouth workshop initiated by John McCarthy. Since then, the field has seen the emergence of many important areas such as knowledge acquisition, knowledge-representation, problem solving, machine learning, natural language processing, perception, robotics, etc. The fruitful scientific investigations in the broad field of AI have also helped in shaping the minds of a large pool of eminent scientists.

Nevertheless, a curious question still remains open: Should Artificial Intelligence be a sub-set or a super-set of computerizable human intelligence? In other words, should a machine such as a robot simply imitate human-like intelligence, or could it instead develop its own way of thinking, learning and acting? If the answer is yes, then what should the new definition of Artificial Intelligence be, or more precisely the definition of Robot Intelligence (if a computer requires a body to act rationally or purposefully)?

2.3. Robot (or autonomous artificial) intelligence

Artificial intelligence literally means intelligence residing in an artificial body such as a robot. However, the content discussed under artificial intelligence suggests that it actually deals with computerizing human intelligence.

As a result, today’s computational tasks underlying computerized human intelligence are dictated by the results of programming undertaken by human beings. None of today’s computers have even mastered a single language of its own or a natural language. Because of this deficiency, a computer has not yet had the innate ability to synthesize human-like intelligence.

In (humanoid) robotics, a computer is treated as the brain of a robot. Hopefully, one day, a robot with a brain could directly develop its own programs, which mimic human intelligence. Therefore, there is plenty of room for the development of a new paradigm of AI, which we can call Robot Intelligence (RI) (or Autonomous Artificial Intelligence (A²I)). RI, or A²I, aims at equipping machines, such as humanoid robots, with the integrated ability to autonomously learn, understand, create, and communicate meanings in the form of images and texts of a natural language.

Then we come to this question: What do we mean by intelligence from an engineering point of view? Refer to Fig. 12. One possible definition of intelligence is as follows:

**Definition 1.** Intelligence is the ability to link perception to actions for the purpose of achieving an intended outcome. Intelligence is a measurable attribute, and is inversely proportional to the effort spent in achieving the intended goal.
Fig. 2. Intelligence is the ability to link perception to actions for the purpose of achieving an intended outcome.

In robotics, we know that:

Motion is an attribute (or physical quantity) engendered by a robot’s mechanism, under the governance of kinematics and dynamics. Kinematics is the study of motion without considering force and torque, while dynamics is the study of motion in relation to force and torque.

In a similar vein, we can say that:

Definition 2. Robot intelligence is an attribute engendered by a robot’s brain, under the governance of causality and rationality. Causality (or causalitics) is the study of intelligence without considering motivation (i.e. value and belief), while rationality (or rationalitics) is the study of intelligence in relation to motivation.

In short, we advocate that intelligence is a function of:

- time,
- richness of internalized physical world,
- richness of internalized conceptual world,
- mental ability of learning,
- mental ability of analysis,
- mental ability of synthesis,
- syntax or habit (i.e. specific patterns of layout, specifically-ordered sequence) of actions,

3. The Robot’s Mental Architecture

It is clear that today’s computational tasks underlying computerized human intelligence are directly developed by humans. This indicates that a computer has not yet had the ability to synthesize human-like intelligence. Maybe, one day, a computer will be able to autonomously develop its own programs which mimic human intelligence. The purpose of this kind of mental development is to make humanoid
robots truly intelligent and autonomous. Clearly, this objective poses a tremendous challenge on how to develop engineering principles, or solutions, which allow to elevate a robot from an intelligent automatic machine to an intelligent autonomous machine.

3.1. The robot’s brain

Because of its human-like shape, a humanoid robot has the potential to exhibit human-like behavior as a result of the interaction of human-like mental processes, which reside in a man-made mental architecture. The human brain has an innate structure, and will remain unchanged if there are no alternations. Then, an immediate question is: What is the innate architecture of a robot’s brain?

We know that today's computers are mostly based on the Von Neumann architecture, which consists of two basic building blocks: (a) the CPU (i.e. central processing unit), and (b) the memory, as shown in Fig. 3(a). All the computational tasks are running via the CPU and memory. In addition, a computer can interface with peripherals or other machines through various I/O systems. Von Neumann has also proposed the concept of stored programs, which will certainly serve as a foundation for the study of knowledge-representation in machines, in contrast to the study of knowledge-representation in humans. We will discuss this topic a little more in a separate section.

In neuroscience, we know that the human brain’s neural system is partitioned into different zones which support concurrent data processing and decision-making. This fact may suggest that a robot’s brain should have a cluster of networked processing units. Accordingly, a robot’s mental architecture should be similar to (if not exactly the same as) that depicted in Fig. 3(b).

Inside a computer’s architecture, the computational tasks run according to algorithms encoded in the form of programs. A computer’s memory serves two

![A Computer's Architecture](image1)

![A Robot's Mental Architecture](image2)

Fig. 3. A robot brain’s architecture and its comparison to the Von Neumann architecture: There are four basic building blocks: (a) the processing unit, (b) the memory unit, (c) the soft agents (e.g. computational tasks), and (d) the input/output unit.
purposes: (a) to store programs, and (b) to store both short-term and long-term data. Two questions arise:

- what should the mental processes in a robot’s brain be?
- what should the content in a robot brain’s memory be?

As a process must operate on content such as input and output, it is necessary to have a clear answer to the second question before one can attempt to partly answer the first.

3.2. The robot’s mental world

It appears obvious that all of us understand what we mean by the real world and the mental world. Hence, it may be unnecessary to define these terms. However, whatever human beings understand does not necessarily imply that they are equally understood by robots or other artificial systems. Therefore, it is necessary to clearly define terms and concepts before any inventive steps can be taken. On the other hand, it is always helpful to bound the meaning of terms and concepts for three purposes:

- to prevent confusing usage of vague terms or phrases (e.g. by artificial intelligence, do we mean computerized human intelligence? By natural language understanding, do we mean the understanding of free texts in a natural language by humans, or by robots?) (Note: in engineering, people never confuse a programming language with the programs which are the results of programming.);
- to cultivate the practice of accurate thinking, reasoning and communicating (e.g. what is a behavior? What is a process? What is a language?);
- to make terms and concepts programmable using a computer language (e.g. can a natural language be represented using a programming language such as C or C++? Can a robot learn a natural language? Can a robot analyze free texts, or documents, of a natural language? Can a robot synthesize free texts, or documents, of a natural language?).

3.2.1. Definition of real world and meaning

We know that the real world encompasses both nature-made and man-made things, which exist in space and time. The presence of things in space, which results in certain spatial arrangement, is generally called configuration in engineering or situation in linguistics. Most importantly, things rarely exist in isolation. Instead, they act upon and interact with one another in space and time. The phenomenon of action-reaction implies that constraints do exist, which govern not only the internal properties of an individual thing (e.g. the ability to act), but also the interaction and interrelation among things (e.g. the ability to react).
It is also the phenomenon of action-reaction which results in the unceasing chain of cause and effect, which is well understood in human history. And, this cause–effect phenomenon gives rise to the notion of *causality* in nature or *logics* in linguistics.

We all agree that action is the source of results (or outcomes). In general, a specific outcome is only achievable by a sequence of ordered actions (or operations). This engenders the notion of *behavior* (or *process*). By definition, behavior (or process) is a sequence of ordered actions (or operations) which result in an intended, or unintended, outcome. Accordingly, the actor of a behavior (or process) is called the *agent* (e.g. a human being or a robot). Since an actor can exhibit behaviors and is situated in a configuration (or situation), the dynamics (or behavior) of a configuration creates the notion of what we generally mean by an *event*. When a set of events act, interact and evolve over a certain range in space and a certain (long) period in time, this phenomenon is referred to as the notion of *episode*. Hence, a real world does have a set of constituent elements (or parts) such as:

\[
\text{entities} \rightarrow \text{constraints} \rightarrow \text{behaviors} \rightarrow \text{events} \rightarrow \text{episodes},
\]

as shown in Fig. 4.

As these constituent elements (or parts) gain their existence due to nature, we call them the constituents of the *physical world*, which simply reveals one aspect of the *real world*. The other aspect of the real world is the so-called *conceptual world*, which can be explained as follows:

Among nature-made creatures, human beings are very special in that they know how to observe results, do analysis, plan actions, and make new things. This type of behavior is normally called *experiments* (or *games* in layman terms). Moreover, the actions involved in an experiment are encapsulated in the notion of *process*, as follows:

\[
\text{input} \rightarrow \text{process} \rightarrow \text{output}.
\]

![Fig. 4. An illustration of the duality of the real world, in which the physical and conceptual worlds co-exist. The properties and causalities behind the physical and conceptual worlds form the body of meanings or knowledge.](image-url)
For example, an experiment can be as simple as using our body to sense the ambient temperature in winter or summer.

Interestingly enough, it is the activity of undertaking experiments, which results in the notion of understanding, which is characterized by another important notion of meaning, as follows:

- the relationship between input and output, which forms part of meaning (i.e. the aspect of constraints);
- the properties of input, which also form part of meaning;
- the properties of output, which are also part of meaning.

The meaning (i.e. properties and constraints) of operations inside a process results in the notion of concept. For example, in manufacturing, we have the concept of heating, solidification, material removal, insertion, etc. And, the accumulation of concepts relevant to the same process gives rise to the notion of topic, such as welding, machining, assembly, packaging, kinematics, dynamics, control, etc.

As a meaning is stable and repeatable, this phenomenon creates the desire, in the mind of a human actor, of communicating the understanding of meanings to others. This desire may be due to the sense of pride (e.g. claiming to be the first one to do such a thing), or the sense of social responsibility (e.g. sharing results among others for the benefit of everyone or a better chance of survival). It seems that the desire to communicate meanings among human beings was the engine behind the invention of natural languages, which may be spoken, written, or both.

A natural language is a man-made thing, which is also a result of evolution and perfection over a long period of time. A natural language serves the purpose of communicating meanings and concepts (or topics) in both spoken and written forms (i.e. a set of symbols and phonemes). It is understood now that the basic building blocks of a natural language are: (a) lexicon (i.e. vocabulary or a set of words), and (b) syntax (i.e. rules which govern the arrangement of words in order to form meaningful phrases and sentences). Furthermore, a construct of, or result of using, a natural language is called a text, which can be hierarchically organized into parts, chapters, sections, paragraphs, sentences, and phrases. Words are the basic elements in a natural language, each of which has: (a) a label (e.g. a symbol in Chinese or a string of characters in Western languages), (b) a sense (i.e. meaning) and (c) a referent (i.e. a corresponding thing in the physical world).

In linguistics, the meaning expressed by a text is called semantics, which consists of both lexical semantics (i.e. the meaning of words and phrases) and logical semantics (i.e. the meaning of sentences and texts). Hence, it becomes clear that a text in a natural language is a vehicle which carries meanings that also reflect logics or causality in the physical world.

When the meanings underlying concepts and topics are encoded into texts, this results in the so-called conceptual world, which do have a physical existence (e.g. textbooks in library). Hence, the conceptual world consists of the following
constituent elements:

\[
\text{lexicon} \rightarrow \text{syntax} \rightarrow \text{logics} \rightarrow \text{concepts} \rightarrow \text{topics},
\]

as shown in Fig. 4.

It goes without saying that a natural language is a tool which serves the purpose of effectively representing concepts and topics in the form of texts. As the accumulated collection of topics forms what we call knowledge, it seems obvious that knowledge is naturally represented by texts in one or more natural languages. (Note: A set of topics relevant to an event or episode forms the notion of a subject such as mathematics, physics, chemistry, etc.) In line with this thinking, the use of knowledge raises the following important issues:

- how do we or robots acquire natural languages which represent knowledge?
- how do we or robots analyze texts of a natural language in order to derive the meanings?
- how do we or robots synthesize texts in a natural language, which convey intended, or understood, meanings (i.e. the issue of knowledge or meaning representation in using texts of a natural language)?

Despite the tremendous number of published works in the literature, we have not yet reached definite, and thorough, answers to the above three questions.\(^{20,21}\)

3.2.2. Definition of the mental world

Research on the understanding of texts in a natural language by humans reveals the existence of mental models.\(^{25}\) The philosopher Kenneth Craik first proposed the term mental model in 1943, and defined them as “small-scale models of reality, which can be used to anticipate events.” Continuous scientific investigation on natural text (or language) understanding by humans further reveals the phenomenon of semantic, episodic and text memories in the human brain.\(^9\)

In engineering terms, these early findings simply suggest that both the physical and conceptual worlds of the real world can partly be perceived by humans, and also be stored in human brain’s memory. This leads us to define the notion of the mental world as the internal representation of the real world, in the form of the internalized physical world and internalized conceptual world, as shown in Fig. 5.

Based on the illustration in Fig. 5, we can further define the notion of an idea and thought as the manifestation of behaviors, events or episodes occurring in the internalized physical world. Since an agent, such as a human being or robot, can perceive and also be perceived, it is legitimate to make a claim such as: “mind in the body; body in the mind.”\(^{11}\)

Now, a consensus seems to be that texts and images do co-exist in the human brain’s memory. For example, when person A says “I have an idea,” it may mean that person A has a visual form of an event, process or episode in his memory. If person B sends to person A the request “Please send me an email which explains
Fig. 5. An illustration of the mental world, which is simply an internalized version of the real world by an agent.

your idea,” person A will certainly be able to synthesize text in a natural language, which conveys the idea in text form to person B. Another example shows that if a car is red, both the car’s image and the concept “red” are stored in the memory of the car’s owner, as a color is a “live” sensation which can only be experienced, but not stored nor visualized inside the human brain.

3.2.3. Generic mental processes

The supremacy of humans over animals is largely a result of our ability to understand (i.e. analyze), create (i.e. synthesize), and experience (i.e. learn). This observation suggests that the innate mechanism underlying mental processes should be based on the interplay among learning, analysis and synthesis, as outlined in Fig. 6.

In this framework, learning plays a key role as a result of the following facts:

- knowledgeable persons perform problem-solving (i.e. analysis) better than non-experts;
- resourceful persons demonstrate a higher level of creativity (i.e. synthesis) than non-experts;
- educated persons know, and can do, more things than those who did not receive any education or training.

Learning is an important skill for humans (or robots) to acquire knowledge and skills. The comprehension of knowledge and enhancement of thinking ability (e.g. analysis and synthesis) will, in turn, make learning easier and more effective.

It is worth noting that the study of artificial intelligence historically focuses on problem-solving (i.e. analysis) and learning.\textsuperscript{14,17} The difficult issue of synthesis (i.e. creativity) has been less studied. In robotics,\textsuperscript{33} it is well understood that the ability to synthesize motion is the key towards the development of autonomous robots. As shown in Fig. 7, if a robot could autonomously synthesize desired motions for
Fig. 6. An illustration of the interplay among learning, analysis and synthesis underlying mental processes in humans or robots: learning plays the central role as the mental world is primarily acquired and enriched through learning and real-time interaction with environment. In addition, analysis and synthesis skills are also acquired in learning.

Fig. 7. Importance of synthesis in developing autonomous robots. An automated motion requires the specification of the desired motion, which is usually given by a human operator. If an agent such as a robot could automatically synthesize its own intended motion, it would become autonomous at the level of motion execution.
the achievement of a given goal as input, then it does not require any human intervention in manually undertaking the planning of desired tasks, actions and motions. Hence, motion synthesis is a necessary condition for an automated robot to become an autonomous robot.

3.3. From meanings to autonomous actions

Today’s robots are automated machines, which have well-developed physical abilities to perform actions. However, the automatic execution of motion requires programming and re-programming, which is manually done by humans. For a robot to gain a certain level of autonomy, one solution is to develop its mental abilities in planning motions, actions or tasks. The purpose of planning motions, actions and tasks is to autonomously specify the desired motions for automatic motion control loops. Without the ability to self-generate desired action sequences, a robot will never become autonomous.

Refer to Fig. 8. An autonomous agent should only receive meanings from (a) a master agent or (b) its own perception system. In this way, there is no need for human intervention in helping the agent do programming or re-programming.

![Fig. 8. From meaning to autonomous action: An agent could derive meanings from (a) visual perception and (b) conversational instructions from a proactive agent. For an agent to gain a minimum level of autonomy, it is necessary for it to self-generate actions from the perceived meanings. With this ability in place, there is no need for a proactive agent to program a reactive agent in using a computer’s programming language.](image-url)
In order to achieve the goal of self-generating action sequences from perceived meanings, two necessary and sufficient conditions must be met, namely:

**Condition 1.** An agent must have the mental ability to derive meanings from texts or images.

**Condition 2.** An agent must have the mental ability to self-generate action sequences in both conceptual and physical worlds.

In the following sections, we are going to address the issues of how to derive meanings from texts and images. As for action sequence generation in the conceptual world, there are many established algorithms to automatic planning of motion sequences in the literature.\(^3\)\(^,\)\(^5\)

### 4. Natural Text Understanding (NTU)

We are living in a meaningful world. The meanings can all be conceptually written into texts of natural languages. And, texts encapsulate logics, concepts and topics, which are the constituents of knowledge. Hence, it is clear that knowledge is represented by using the tool called *natural language*. Human beings are skillful in manipulating knowledge in various ways such as:

- knowledge acquisition (i.e. learning and analysis);
- knowledge production (i.e. learning and synthesis);
- knowledge processing (i.e. filtering and transformation);
- knowledge communication (in a written or spoken language).

The manipulation of knowledge depends on the use of a natural language.

#### 4.1. *A generic application scenario*

A great deal has been understood about how humans intelligently manipulate knowledge and computerize human skills in learning and analysis (i.e. problem-solving) with a programming language such as Prolog or Common Lisp.\(^14\)\(^,\)\(^19\) However, it is still an open issue as to how to make a humanoid robot understand meanings or knowledge in the form of texts of a natural language.

In computer vision, we know that an image can have many transformed representations of its original form (i.e. a matrix of pixels), and that an image transform serves the purpose of facilitating feature extraction, but it does not directly lend itself to image understanding. Similarly, the transformation of texts into other forms of symbolic representation only serves the purpose of text processing, and will not directly aid in text understanding by a robot or computer.

When a meaning is understood by an agent, this agent should be able to explain the meaning in various forms, such as the text of a natural language, gestures, or appropriate behaviors. This prerequisite suggests that text understanding should
invoke a synthesis module to validate, verify or demonstrate the meaning understood in one way or another.

Therefore, the generic application scenario of NTU is better understood with the illustration of two agents engaged in conversation, as shown in Fig. 9.

There are three possible outcomes as a result of the interaction between two agents using a natural language:

(i) *Deny*: Speaker receives the utterance from Listener, who contradicts the composition of meanings. In this case, Listener is most likely an expert and Speaker is a learner.

(ii) *Optimize*: Speaker receives the utterance from Listener, who suggests a better composition of meanings. In this case, Listener is most likely an expert and Speaker is a learner.

(iii) *Confirm*: Speaker receives the utterance from Listener, who asks for clarification. In this case, Listener is most likely a learner.

If Listener is an expert, both outcomes of Deny and Optimize are possible. In this case, Speaker behaves like a learner. If Listener is not an expert, the most likely outcome will be Confirm. In this case, Listener behaves like a learner.
If the outcome is Deny, Listener contradicts Speaker’s statement and Speaker learns from Listener. If the outcome is Optimize, Listener suggests a better way of representing the meaning or statement verbally as well as visually. (It is similar to the Auto-Correct feature of MS-Word, which gives suggestions, when a user commits a typo or a grammatical error.) If the outcome is Confirm, it indicates that Listener has difficulties in grasping the intention of Speaker. As a result, Listener further clarifies with Speaker.

4.2. Limitations of existing research of NTU

Natural Text Understanding (NTU) is important for developing any natural language based applications. The meaning of a word should consist of sense (i.e. relationship between words) and reference (i.e. corresponding thing in the physical world). But, all the existing NTU systems consider only the sense of the word in order to derive the symbolic description of a sentence, which may not prove the understanding of meanings. In other words, all the NTU systems transform the texts into other forms of symbolic representation. It serves the purpose of text processing, and will not directly aid in the understanding of meanings by a robot or computer.

Currently, there are many computerized dictionaries and lexicon databases commercially available. Although they manipulate a lexicon and syntax, what they do exactly is the retrieval of annotations which match a request in the form of a word as input. These annotations are often encoded in the form of strings of characters. Sometimes, they may have visual annotations (i.e. images) which are pre-programmed by humans. However, they cannot compose a text of their own in order to convey an intended meaning.

4.3. A meaning-centric framework

An NTU system should map the free text to its meaning corresponding to the physical world (i.e. visualization of meanings of a natural text) as well as conceptual world (i.e. synthesis of free text), instead of transforming free texts into other forms of symbolic representation. What we exactly mean by understanding the meanings of a text by a robot are as follows:

- The robot should learn the elementary meanings of words: When a robot knows the properties and constraints of a word such as lexical constraints, syntactic constraints, semantic constraints, and physical constraints (a later section gives more explanation to these constraints), we can say that the robot understands the elementary meaning of the word. We call this process the learning of elementary meanings.
- The robot should master a natural language: A natural language imposes its own syntactical structures and grammatical rules on the lexicon. Through the application of syntactical structures and grammatical rules, sentences and texts can
be constructed out of words. This is a prerequisite for the synthesis of composite meanings in the form of sentences and texts by humans or robots.

- The robot should understand composite meanings of texts: A text consists of strings of words, which are arranged according to the syntactical structures and grammatical rules of a natural language. Based on the understanding of the elementary meanings of words and the recognition of syntactical structures underlying a text, a robot should be able to interpret the composite meanings. Subsequently, it should be able to visualize the understood meanings. We call this process the analysis of composite meanings.

As the synthesis skill is primarily based on the manipulation of internalized knowledge by an agent, this implies that an intelligent agent must have accumulated a certain amount of knowledge through the process of learning. Clearly, the mental processes underlying natural text understanding by intelligent robots such as humanoid robots must include learning, analysis and synthesis as shown in Fig. 10.

Fig. 10. A meaning-centric framework: A robot must build a mental world through learning before it is ready to acquire skills for analysis and synthesis through learning as well. The interaction of a robot with its own environment will allow a robot to further enrich its mental world, which is not a static database. With the mental world and the skills for analysis and synthesis in place, a robot should be able to interpret an input text and also to synthesize an output text which conveys an intended meaning.
The primary steps involved in the meaning-centric framework include the following:

(i) Through learning, a robot perceives and internalizes the meaning of physical and conceptual worlds in order to form its mental world (or mind). This process builds up the internalized knowledge in the robot’s mind or mental world. And, a robot’s performance in undertaking analysis and synthesis depends on the enrichment of its mental world (or internalized knowledge). This explains why an educated robot should perform better than an uneducated robot. Thus, a robot should enrich its mental world before it becomes capable of undertaking analysis and synthesis. And, learning is an important process behind the enrichment of a robot’s mental world.

(ii) During analysis, a robot receives an input text from an agent in the real world, and derives the text’s meaning by referencing to the internalized knowledge in the mental world (or understands the text’s meaning in its own way).

(iii) During synthesis, a robot composes its own text based on the interpreted, or intended, meaning. The understood, or intended, meaning can be conveyed not only in the form of text but also in the form of gesture, scenery representation, or appropriate behaviors, etc.

4.3.1. **Text acquisition by robots**

A text can be acquired in various ways such as input from a keyboard, input from a speech recognition system, and input from the scanning of a printed document. Although it is still necessary to improve speech recognition, we restrict our discussions to the difficulties encountered in analyzing, synthesizing, and learning free texts of a natural language such as English.

4.3.2. **Text analysis by robots**

Refer to Fig. 10. Text analysis invokes a number of sequential modules. The most important ones are:

- **Semantics Inferring**: The purpose of this module is to derive the meaning of words and sentences in an input text. It is well-understood that a standard approach to this problem is to apply the following procedure:
  
  Step 1: Segmentation of an input text into words, phrases and sentences. Do filtering in order to eliminate errors (e.g. misspellings), if any.
  
  Step 2: Classification of words into nouns, pronouns, verbs, adverbs, auxiliary verbs, adjectives, prepositions, quantifiers, complementizers, and determiners.
  
  Step 3: Identification of syntactical structures, in terms of the noun phrases (NP) and verb phrases (VP), underlying the sentences.
  
  Step 4: Determination of the meanings underlying words.
Step 5: Determination of the meanings underlying phrases.
Step 6: Determination of the meanings and logics underlying sentences.

- **Concept Inferring**: A concept refers to the meaning of an event (i.e. a set of ordered interactions or behaviors occurring in space and time) in a physical world. Its representation in the conceptual world is in the form of one, or more paragraphs in a text. Hence, the aim of this module is to identify the conceptual grouping of paragraphs, given the meanings and logics of sentences as input.

- **Topic Inferring**: A topic refers to the meaning of an episode (i.e. a set of ordered events occurring in space and time) in a physical world. A topic consists of a set of coherent concepts, which are arranged together according to certain logics. Therefore, the task of topic inferring is to identify the conceptual grouping of concepts into topics, which have distinct meanings and logics.

It is clear that topic inferring depends on concept inferring which in turn is largely compromised by the success of semantics inferring. A close examination of the standard procedure underlying semantics inferring reveals that it is actually doing two things: (a) text processing (Steps 1–3), and (b) meaning determination (Steps 4–6). If these two operations are to be autonomously undertaken by an agent such as a robot or computer (i.e. without human intervention), then the following necessary condition must be met:

**Necessary Condition 1.** An agent has an internalized conceptual world in which the lexicon and syntax are well understood by the agent itself.

Hence, the first difficulty in approaching the issue of natural text understanding by a robot is the question of how to make a humanoid robot learn a natural language such as English.

### 4.3.3. Text synthesis by robots

Today’s robots are solely automated machines. The robots of the future will certainly be sociable creatures, which co-exist with human masters. A sociable robot must have the ability to understand the meanings encoded in a spoken or written form of a natural language by its human masters.

Although there are many commercially available computerized dictionary (e.g. iFinger for Oxford English Dictionary) and hybrid dictionary/thesaurus (e.g. WordNet), none of them enable a computer to creatively compose a text of its own. If a computer or robot could understand the lexicon and syntax of a natural language such as English, then there is a chance for it to synthesize a text which conveys the meanings intended by itself, instead of the meanings preprogrammed by humans.
Assume that a robot could understand the meaning of a text by the following procedure:

1. semantics Inferring;
2. concept Inferring;
3. topic Inferring.

Then, it is possible to have it synthesize a similar text, which conveys the understood meaning, by an inverse procedure, as outlined in Fig. 10, which consists of:

- **Topic Formation**: Having a story, or subject, in mind, which is to be communicated to others in a spoken or written form of a natural language, the aim of topic formation is to plan a set of ordered episodes, which could effectively convey the meaning of a story, or subject.

- **Concept Formation**: An episode normally encapsulates a set of ordered events, which occur within their respective configurations. Hence, the purpose of concept formation is to plan a set of ordered events and their configurations, which accurately display the content of an episode.

- **Semantics Formation**: An event can further be decomposed into a set of ordered interactions or behaviors. And, an interaction or behavior could be described by one or more sentences. Therefore, semantics formation normally invokes the following procedure:

  Step 1: Decomposition of an event into a set of interactions or behaviors.
  Step 2: Identifying the agents (e.g. actors and reactors) involved in the interactions or behaviors.
  Step 3: Identifying the constraints imposed on the agents.
  Step 4: Identifying the nouns which match with the agents.
  Step 5: Identifying the verbs which describe the constraints.
  Step 6: Identifying the most appropriate syntactical structure which arranges the nouns and verbs together.
  Step 7: Construct sentence(s) by appropriately filling words into the chosen syntactical structure.

It is clear that the formation of topics and concepts cannot occur in the internalized conceptual world, because a story or subject has no corresponding text yet. Instead, it can only occur in the internalized physical world. As a result, the second difficulty in undertaking natural text understanding by a robot is its ability to visualize topics and concepts in the internalized physical world. In other words, an agent having the ability to synthesize texts of a natural language must meet the following necessary condition:

**Necessary Condition 2.** An agent has an internalized physical world, and could undertake imagination (i.e. autonomous simulation) inside it.
4.3.4. Learning by robots

A free text, which is the input to one agent, can be viewed as a “program” which is the result of linguistic programming (i.e. programming with a natural language) done by another agent. Hence, if an agent is able to understand a free text, it is necessary for itself to be programmable or reprogrammable at a linguistic level. Without an internalized conceptual world, a robot’s brain will not be able to undertake linguistic programming; without the ability to undertake linguistic programming, a robot will not become a sociable robot.

Admittedly, a sociable or educable robot must be reprogrammable at a linguistic level. However, we are still far away from reaching this goal. Nevertheless, we acknowledge the tremendous effort spent on building up electronic version of knowledge bases such as WordNet (Princeton University) and MindNet (Microsoft), which help address the issue of constructing internalized conceptual world (i.e. necessary condition 1). However, solutions to the issue of constructing internalized physical world (i.e. necessary condition 2) are expected to emerge soon or later.

Accordingly, the third difficulty in undertaking natural text understanding by a robot is the requirement of being programmable or reprogrammable at a linguistic level. That is:

Necessary Condition 3. An agent is linguistically programmable or reprogrammable in order to enrich its mental world.

Natural language is not an innate ability of any creature on earth. It must be learnt in one way or another. Refer to Fig. 11. Learning is a process which maps external phenomenon manifested in the form of observable features and states into an internalized representation in the form of descriptive features and states.

In general, learning must invoke four functional modules: (a) modelling of properties and constraints, (b) optimization, (c) durable representation of properties and constraints, and (d) evaluation/experiment.

In engineering, there are many well-developed solutions to model the properties and constraints of continuous-time, or discrete-time, systems and processes. In the physical world, the constraints governing the dynamic behaviors of many physical systems can be described by either the continuous-time state-space model:

\[
\begin{align*}
\dot{X}(t) &= A \cdot X(t) + B \cdot u(t), \\
Y(t) &= C \cdot X(t) + D \cdot u(t),
\end{align*}
\] (1)

or the discrete-time state-space model:

\[
\begin{align*}
X(k + 1) &= A \cdot X(k) + B \cdot u(k), \\
Y(k) &= C \cdot X(k) + D \cdot u(k),
\end{align*}
\] (2)

where \(X(\cdot)\) is the system’s state vector, \(u(\cdot)\) the input vector or scalar, and \(Y(\cdot)\) the output vector or scalar.

In the above equations, \((A, B, C, D)\) are coefficient matrices or vectors which describe the properties of a system under learning. These matrices also define the
Fig. 11. Robot learning consists of modelling, optimization and representation, which are complemented by evaluation/experiment. The outcome of learning should be the internalized representation understood by an agent. Depending on the nature of a learning process, we have: (a) unsupervised learning, which only invokes modelling, optimization and representation, (b) reinforcement learning, which additionally includes self-evaluation/experiment, and (c) supervised learning, which maps externally imposed representation (or sample) into an internalized representation.

so-called feature space, and are usually estimated from sample observations of a system’s input, output and internal states. This process of estimating \((A, B, C, D)\) is called optimization. The outcome of optimization is the durable representation of a system under learning.

In the physical world, all systems or processes are statistical in nature because of uncertainty, noise, and disturbance. If the size of sample observations must be large enough in order to make optimization significant, statistical inference must be invoked. In Eq. (2), we can omit \((B, D)\) (i.e. set them to zero). If we further treat \(A\) as a matrix of transition probabilities and \(C\) as a matrix of emission (or observation) probabilities, the discrete-time state-space model will become the famous Markov Model proposed by Andrei A. Markov in the early 1910s. If we introduce the probabilities of initial system states, denoted as \(\Pi\), then we obtain the so-called Hidden Markov Model (HMM).

Refer to Eq. (2) again. The system’s state vector \(X(\cdot)\) is composed of a set of internal state variables \((x_1, x_2, \ldots, x_n)\). If the system under learning is stochastic in nature, each state variable can be treated as a random variable. In this case, the square matrix \(A\) becomes an interaction matrix if \(A\) is not a diagonal matrix.
Interestingly enough, if we consider $A$ to be time-varying, $A(k)$ will describe the interaction among the internal states $(x_1, x_2, \ldots, x_n)$ at time instant $k$. The outcomes of this interaction give rise to the values of the internal states $(x_1, x_2, \ldots, x_n)$ at time instant $k + 1$. Graphically, $A(k)$ forms one layer of mapping from $(x_1, x_2, \ldots, x_n)$ at time instant $k$ to $(x_1, x_2, \ldots, x_n)$ at time instant $k + 1$. If we stack a series of successive layers together, what we obtain is exactly the internal layers of a statistical neural network. Furthermore, the output layer can be readily formed by the mapping governed by $Y(k) = C(k) \cdot X(k)$, whereas the input layer is determined by the mapping governed by $B(k) \cdot u(k)$.

Therefore, we would like to claim that the state-space model, HMM and neural network could be treated as the inter-related and innate principle underlying the modelling of properties and constraints governing the dynamic behaviors of most systems on earth.

Now we return to the question of how to model a natural language. In general, a natural language modelling must consider these three issues:

- how to model and represent elementary meanings of words;
- how to model and represent the transition, interaction and prediction of words in the conceptual world;\textsuperscript{31,32}
- how to model and represent the transition, interaction and prediction of words in the physical world in order to derive composite meanings.

In the literature, a great deal of effort is narrowly focused on the second issue. Little has been done with regard to the first and third issues. It is our belief that what is crucial in a natural language modelling is not only the interaction among words but also the essence of meanings. This highlights the difference between our research and existing works in the literature.

In view of the large number of existing natural languages, and the complexity of their lexicons, it may be wise to adopt a developmental approach to incrementally acquiring and organizing the lexicon of a natural language by a robot. In this way, a relevant question arises: What is the innate mechanism (or methodology) for a robot to autonomously acquire and organize the lexicon of a natural language? With this question in mind, we are obviously no longer concerned about how to computerize a printed dictionary. As the lexicon serves as the vehicle to support both natural text understanding and free text synthesis, there is the question of whether a lexicon internalized by a robot should incorporate lexical constraints together with syntactic and physical constraints.

Here, we propose an active memory model called WordDNA to represent the generic elementary meanings (i.e. properties and constraints) of words in both physical and conceptual worlds as illustrated in Fig. 12.

WordDNA can be implemented with an object-oriented programming language so that it can serve as a base class which can be Duplicated, Nurtured and Accomplished (DNA).
Fig. 12. Active memory model to represent properties and constraints (i.e., meanings) associated to words in both physical and conceptual worlds. With this model in place, the meanings of a word can be incrementally acquired, explained and visualized. In addition, a set of such words could be put together to interact so as to engender composite meanings.
4.4. **Implementation**

As discussed above, the proposed meaning-centric framework for NTU by robots consists of the following major processes:

- modelling and learning of elementary meanings;
- analysis and synthesis of elementary meanings (i.e. meanings underlying words);
- analysis and synthesis of composite meanings (i.e. meanings underlying sentences).

In order to effectively implement the proposed meaning-centric framework for NTU by robots, we have started the development of a software agent called *KnowNet*.

4.4.1. **Software agent: KnowNet**

*KnowNet* is an on-line learning and tutoring system for the English language, which supports bi-directional learning activities such as human learning from robots and robot learning from humans.

*KnowNet* is based on the above-mentioned active memory model called *WordDNA*, which is implemented in Java (an object-oriented programming technique). In this way, *KnowNet* has an internalized organization (i.e. representation) of a lexicon, which could be incrementally acquired by a robot through real-time interaction with its environment (including human masters). The base class *WordDNA* in Java of the proposed active memory model has the following fields as shown in Fig. 12: (a) properties in the physical world, (b) constraints in the physical world, (c) properties in the conceptual world, and (d) constraints in the conceptual world.

The proposed active memory model enables a robot to acquire not only lexical and syntactical constraints/properties associated to words, but also their corresponding physical constraints/properties. In addition, the active memory measures the degree of understanding of learnt words by a robot, which could be incrementally enhanced through its real-time interaction with a dynamically changing environment.

4.4.2. **Learning of elementary meanings in KnowNet**

On top of annotations associated to a word, an important aspect of elementary meaning acquisition is the incremental learning of various constraints/properties in both physical and conceptual worlds. So far, *KnowNet* supports the interactive learning of four possible types of constraints underlying a word:

- lexical constraints;
- syntactical constraints;
- semantic constraints;
- physical constraints.
4.4.3. Analysis and synthesis of elementary meanings in KnowNet

When the “explain” or “visualize” method of a word’s object is invoked, KnowNet performs the visualization of the elementary meaning of the word. And, it also composes its own text to convey the understood meaning, as the example shown in Fig. 13.

4.4.4. Analysis and synthesis of composite meanings in KnowNet

When a sentence is given as input to the text understanding system, the objects corresponding to the words of this sentence interact with each other in order to analyze the input text and also to synthesize a text as output.

For example, when the sentence “the rabbit eats the tiger” is given as input text, the meaning of this sentence is first interpreted in the analysis stage. The robot conveys its understood meaning in the form of text and images. By applying physical constraints, it reasons that the rabbit cannot eat the tiger. It suggests two possible alternatives:

- Rule of substitution: “The rabbit eats something which is vegetable.”
- Rule of alternation: “The tiger eats the rabbit.”

Fig. 13. Graphical User Interface of KnowNet. When an user keys in the request about the word “car,” KnowNet will explain the meanings in both physical and conceptual worlds.
5. Natural Scene Understanding (NSU) by Robots

With regard to the incremental learning of the mental world, it is necessary for a robot to autonomously acquire physical models of conceptual objects. This task depends on the mental abilities of interpreting images. Therefore, the ability to understanding texts is strongly coupled with the ability to understanding images. In engineering research, these two topics are investigated by separate research groups. It is our opinion that these two topics should be studied together under a similar meaning-centric framework.

Therefore, A question posed in this section is: “Can machines see?” Although technologies have advanced greatly compared to a half century ago, research in machine vision still has numerous unresolved issues. This failure of intelligent interpretation of images is mainly due to the absence of a realistic framework which encompasses learning, analysis and synthesis.

5.1. Problematic aspects of equating vision with description

It is recognized that there are five main problematic areas faced by machine vision in understanding images or scenes. They are:

(i) the inherent ineffability of pictures (inability of language to put certain things into words);
(ii) the questionable ontological (over-abundance of possible objects) status of “objects” of which scenes are composed;
(iii) the impossibility of segmenting images (object recognition) in a consistent and principled manner;
(iv) the potential involvement of the entire cognitive system of the perceiver in interpreting image fragments both small and large;
(v) the need for a homunculus implied in postulating a language-like format for the ultimate stage of visual representation.

5.2. Meaning-centric framework for NSU by robots

In order to address the above challenges faced by machine vision, we propose a meaning-centric framework which remedies two important failures in the classical framework of image understanding, namely:

• the negligence of event and episodic recognition;
• the missing link between scene (or image) analysis and text synthesis (i.e. scene description).

Refer to Fig. 14. NSU is a mental process for the purpose of deriving description (i.e. text) from the input of images. This mental process consists of three basic modules, namely: learning, scene analysis and scene description (i.e. text synthesis) in which a ‘lively’ acquired image is semantically interpreted at three hierarchical
levels for both analysis (the recognition of object, event and episode) and synthesis (the planning of topics, concepts and semantics).

5.3. Natural scene analysis

Although image segmentation and feature grouping are still two major obstacles faced by machine vision, the successful interpretation of a scene or image must start with scene analysis, which must be built on three hierarchical levels of recognition, namely:

- **Object Recognition**: Object recognition refers to grouping the visual features belonging to a single physical object, and assigning a “sense” to it.

- **Event Recognition**: It refers to the understanding of relationship and behaviors among recognized objects. Event recognition will not only take into account entities (i.e. senses) associated with individual objects but also configurations and constraints among these objects.

- **Episode Recognition**: Episodic Recognition refers to the “story-like” combination of interrelated events evolving in space and time. This step is formulated to allow “documentation” of actions (events) that evolve in time. Hence, the image-to-text synthesis can be achieved and visual learning of the real environment is made possible. The capture of changes or occurrences evolving in time is undeniably critical to visual learning.
5.4. **Natural scene description**

Scene description refers to the generation of texts based on the meanings derived from scene analysis. The procedure of undertaking scene description is the same as that for text synthesis as outlined in Sec. 4.3.3.

5.5. **Natural scene learning**

The advances in computer graphics have generated a large number of numerical methods to model geometrical objects and behaviors. It is no longer difficult to synthesize realistic nature-like scenes.

However, it is an open issue as to how to autonomously learn visual objects and their behaviors by robots themselves. At this point in time, there is still no definite answer.

6. **Concluding Remarks**

This paper gives a high level description of our research effort on two aspects behind the development of machines or robots that understand meanings in the form of texts and images. We acknowledge that the paper offers a general assessment of some difficult issues underlying the mental development of intelligent machines or robots, which aim to possess the ability of natural language learning and natural text understanding.

We also acknowledge that the increase in complexity of a humanoid robot will give rise to the advantage of having complex, or even human-like, behaviors which otherwise are not possible with a simple arm manipulator. Most importantly, the increase in organized complexity of humanoid robots will certainly bring up many challenging research issues which are still looking for engineering solutions. At a functional level, these challenging issues include:

- **Learning Machine**: A robot can be preprogrammed to repetitively execute instructions which solely reflect the intention of its human masters. This mode of operation will not make a robot adaptable to a dynamically changing environment where unpredictable events may occur in both space and time. On the other hand, without a learning capability, a robot will never be able to acquire new skills or behaviors in an autonomous manner.

- **Thinking Machine**: Learning includes, but is much more than, memorizing. Most importantly, a learning machine must have the analytical skill in undertaking analysis and synthesis, which invoke the manipulation of conceptual entities (or knowledge). In other words, a machine without the innate mechanism of thinking and representation can never internalize any learned experience.

- **Seeing Machine**: We are living in a physical world, which generically consists of entities (or objects), constraints and behaviors (i.e. ordered sequences of actions and interactions among entities). Without vision, a machine would not be able
to perceive a physical world. This shortcoming will tremendously undermine a machine’s ability to act and interact with others. Therefore, a humanoid robot must be a seeing machine.

• **Reading Machine:** Our real world encompasses both natural and man-made entities (or objects). Among the man-made entities, a very important category of human invention is natural languages. It goes without saying that all meanings which are the results of understanding can be described in terms of texts of a written language. This fact also suggests that a natural language is a system for knowledge representation. To a great extent, we can say that all the published texts, which objectively describe meanings, form a *conceptual world*. Since the perception of the conceptual world is equally important as the perception of the physical world by any intelligent creature, a humanoid robot should master at least one natural language.

• **Listening Machine:** When humans and intelligent machines co-exist, it is natural to expect that machines learn and speak human languages, instead of the current situation that humans have to learn and use a computer’s languages, such as a high-level programming language, in order to instruct machines to work for us. Therefore, it is a prerequisite for a humanoid robot to master conversational speech with at least one natural language before it can become a sociable machine.

• **Autonomous Machine:** With the advances achieved in control engineering today, it is not a difficult job to develop an automated machine which repetitively produces a desired outcome, such as motion. As we know, it is relatively easy to specify a desired outcome of a single action, and to realize it with an automatic feedback control loop. However, the challenge is to develop a machine in which the human invention remains at a higher level where the desired outcome, instructed by humans, is not achievable by a single action, but a sequence of ordered actions (e.g. outcomes such as “having won a match,” “having cleaned a house,” etc).

It is true that the scientific inquiry into the question of how the human mind works has resulted in a large body of knowledge in cognition, psychology and learning. However, we acknowledge that the development of intelligent machines or systems is still its infancy, because partly computerized human intelligence does not necessarily mean machine intelligence. With this opinion in mind, it is expected that any initiative in the direction of making robots learn, organize, analyze and synthesize knowledge in the form of texts or images will be most welcome.

On the other hand, we would like to repeat the fact that today’s robots still rely on re-programming with a computer’s programming language. However, it is desirable to make future robots re-programmable at the linguistic level with a natural language. This is a prerequisite toward the development of truly autonomous and sociable robots.

Lastly, the deployment of intelligent machines or robots into human society will logically bring up concerns about their energy consumption, safety, reliability, and
user-friendliness. Admittedly, there is plenty of room for new knowledge, innovation, invention and applications in the research effort towards machines that understand meanings, and also act according to intended or understood meanings.

References