Natural language understanding based on Mental Image Description Language *Lmd* and its application to language-centered robot manipulation

Eri Tauchi, Yuzo Sakuramoto, Tetsushi Oka, Kaoru Sugita, and Masao Yokota

Fukuoka Institute of Technology, 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka-shi, 811-0295, Japan
(Tel : 81-92-606-5897; Fax : 81-92-606-8923)
(yokota@fit.ac.jp)

Abstract: The authors have been working on natural language understanding based on the knowledge representation language L_{md} and its application to robot manipulation by verbal suggestion. The most remarkable feature of L_{md} is its capability of formalizing spatiotemporal events in good correspondence with human/robotic sensations and actions, which can lead to integrated computation of sensory, motory and conceptual information. This paper describes briefly the process from text to robot action via semantic representation in L_{md} and the experimental results of robot manipulation driven by verbal suggestion.

Keywords: Natural language understanding, Robot manipulation, Knowledge representation language.

I. INTRODUCTION

Natural language is the most important among the various information media for ordinary people because it can convey the exact intention of the sender to the receiver due to its syntax and semantics common to its users, which is not necessarily the case for another medium such as gesture. Therefore, natural language can play the most crucial role in intuitive human-robot interaction, namely, interaction between ordinary (or non-expert) people and home robots. This case can be generalized as shown in Fig.1, where robots must comprehend natural language, understand sensory events precisely and plan appropriate actions based on a certain knowledge representation language (KRL) with a good capability of formulating spatiotemporal events corresponding to human/robotic sensations and actions in the real world.

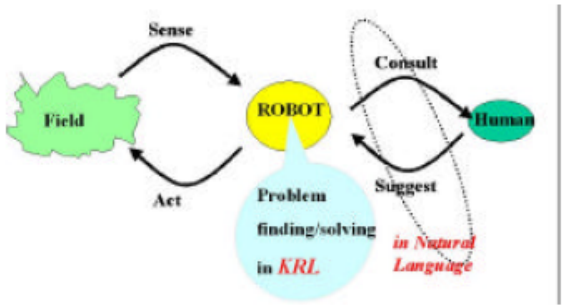

Fig.1. Intuitive human-robot interaction

Conventionally, such quasi-natural language expressions as 'move(10meters)' and so on, uniquely related to computer programs, were employed for deploying sensors/ motors in robotic systems [e.g., 1,2]. These kinds of expressions, however, were very specific to devices and apt to have miscellaneous syntactic variants among them such as *move(10meters, quickly)*, *move(quickly, 10meters, leftward)*, etc. for motors and *find(object, red)*, *find(object, round, red)*, etc. for sensors. This is very inconvenient for communications especially between devices unknown to each other and therefore it is very important to develop such a language as is universal among all kinds of equipments. The language L_{md} [3-5], proposed in Mental Image Directed Semantic Theory (MIDST), was originally for formalizing the natural semantics, that is, the semantics specific to humans, but it is general enough for the artificial semantics, that is, the semantics specific to each artificial device such as robot. The most remarkable feature of L_{md} is its capability of formalizing spatiotemporal events in good correspondence with human/robotic sensations and actions, which can lead to integrated computation of sensory, motory and conceptual information while the other similar knowledge representation languages are designed to describe the logical relations among conceptual primitives represented by natural-language words or formally defined tokens [e.g., 6]. This language has already been implemented on several types of computerized intelligent systems [e.g., 4] and there is a feedback loop between them for their mutual refinement, unlike other similar ones [e.g., 6, 7].

This paper describes briefly the processes from text to robot action via semantic representation in L_{md} [8,9] and the experimental results of robot manipulation by text.

II. NATURAL LANGUAGE UNDERSTANGING

For comprehensible communication with humans, robots must understand natural language *semantically* and *pragmatically* as shown in Figure 2. Semantic understanding means associating symbols to conceptual images of matters (i.e., objects or events), and pragmatic understanding means anchoring symbols to

real matters by unifying conceptual images with perceptual images. Robot manipulation by verbal suggestion here is defined as human-robot interaction where a human presents a robot a verbal suggestion expressing of his/her intention and the robot behaviouralizes its conception, namely, semantic and pragmatic understanding of the suggestion, and that repeatedly. Figure 3 and 4 show the process flows involved.

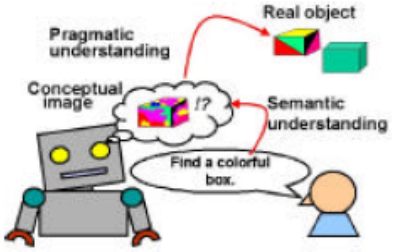

Fig.2. Semantic and pragmatic understanding

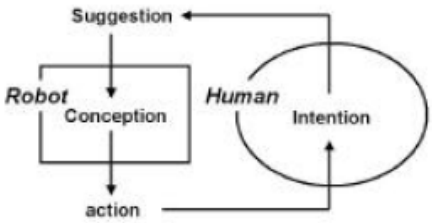

Fig.3. Robot manipulation as human-robot interaction.

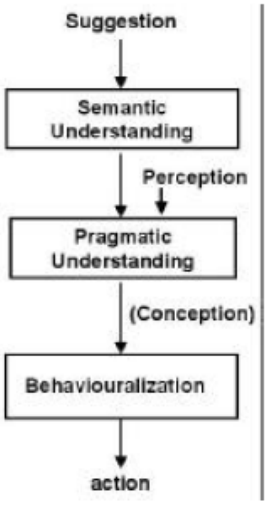

Fig.4. Detailed flow from suggestion to action in robots

The processes in Fig.3 and 4 can be formalized as follows, where the pair of $\mathbf{P_i}$ and $\mathbf{Def_i}$ is called 'Conception' for the i-th action and denoted by $\mathbf{C_i}$.

$\mathbf{Int_i} \Rightarrow \mathbf{T_i}$

$\mathbf{T_i}, \mathbf{K_L} \Rightarrow \mathbf{S_i}$

$\mathbf{S_i}, \mathbf{Per_i}, \mathbf{K_D} \Rightarrow \mathbf{P_i}, \mathbf{Def_i}\ (=\mathbf{C_i})$

$\mathbf{P_i}, \mathbf{Def_i}, \mathbf{K_D} \Rightarrow \mathbf{A_i}$

, where

$\mathbf{Int_i}$: The i-th intention by the human,

$\mathbf{T_i}$: The i-th suggestion by the human,

$\mathbf{S_i}$: Semantic understanding of the i-th suggestion,

$\mathbf{K_L}$: Linguistic knowledge in the robot,

$\mathbf{K_D}$: Domain-specific knowledge in the robot at the i-th session,

$\mathbf{Per_i}$: Perception of the environment in association with the i-th suggestion,

$\mathbf{P_i}$: Pragmatic understanding of the i-th suggestion,

$\mathbf{Def_i}$: Default specification for the i-th action,

$\mathbf{A_i}$: The i-th action by the robot,

$\Rightarrow$: Conversion process (e.g., inference, translation).

1. Semantic understanding

As shown in Figure 6, natural language expression (i.e., surface structure) and L_{ml} expression (i.e., conceptual structure) are mutually translatable through surface dependency structure by utilizing syntactic rules and word meaning descriptions [3].

A word meaning description M_w is given by (1) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$$M_w \Leftrightarrow [C_p:U_p] \quad (1)$$

The C_p of a word W is a logical formula while its U_p is a set of operations for unifying the C_ps of W's syntactic governors or dependents. For example, the meaning of the English verb 'carry' is approximately given by (2), where A_{12} is the attribute of "Physical location".

$$[(\exists x,y,p_1,p_2,k)L(x,x,p_1,p_2,A_{12},G_t,k)\Pi L(x,y,p_1,p_2,A_{12},G_t,k)\wedge x\neq y\wedge p_1\neq p_2: \mathrm{ARG(Dep.1,x)}; \mathrm{ARG(Dep.2,y)};] \quad (2)$$

The U_p above consists of two operations to unify the arguments of the first dependent (Dep.1) and the second dependent (Dep.2) of the current word with the variables x and y, respectively. Here, Dep.1 and Dep.2 refer to the 'subject' and the 'object' of 'carry', respectively. Therefore, the sentence '*Mary carries a book*' is translated into (3).

$$(\exists y,p_1,p_2,k)L(\mathrm{Mary},\mathrm{Mary},p_1,p_2,A_{12},G_t,k)\Pi\ L(\mathrm{Mary},y,p_1,p_2,A_{12},G_t,k)\wedge \mathrm{Mary}\neq y\wedge p_1\neq p_2\wedge \mathrm{book}(y) \quad (3)$$

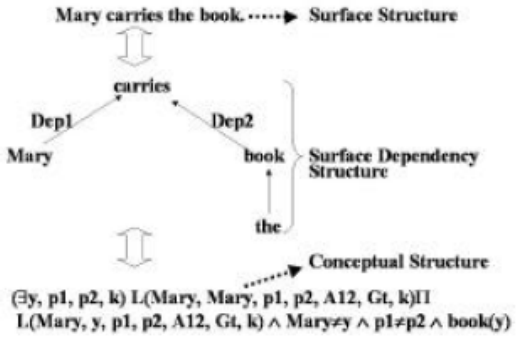

Fig.5. Mutual conversion between text and L_{ml}

For another example, the meaning description of the English preposition 'through' is also approximately given by (4).

$$[(\exists x,y,p_1,z,p_3,g,k,p4,k0)\ (\underline{L(x,y,p1,z,A_{12},g,k)}\bullet L(x,y,z,p_3,A_{12},g,k))\Pi\Pi L(x,y,p_4,p_4,A_{13},g,k_0) \wedge p1\neq z\wedge z\neq p3: \mathrm{ARG(Dep.1,z)}; \mathrm{IF(Gov=Verb)}\rightarrow \mathrm{PAT(Gov,(1,1))}; \mathrm{IF(Gov=Noun)}\rightarrow \mathrm{ARG(Gov,y)};] \quad (4)$$

The U_p above is for unifying the C_ps of the very word,

its governor (Gov, a verb or a noun) and its dependent (Dep.1, a noun). The second argument (1,1) of the command PAT indicates the underlined part of (4) and in general (i,j) refers to the partial formula covering from the ith to the jth atomic formula of the current C_p. This part is the pattern common to both the C_ps to be unified and called 'Unification Handle (U_h)' and when missing, the C_ps are to be combined simply with '∧'.

Therefore the sentences S1-S3 are interpreted as (5)-(7), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (8) is the C_p of the adjective 'long' implying 'there is some value greater than some standard of length (A_{02})' which is often simplified as (8').

(S1) The train runs through the tunnel.

$$(\exists x,y,p_1,z,p_3,k,p_4,k_0)\ (\underline{L(x,y,p_1,z,A_{12},G_t,k)}\bullet L(x,y,z,p_3,A_{12},G_t,k))\Pi L(x,y,p_4,p_4,A_{13},G_t,k_0) \wedge p_1\neq z\wedge z\neq p_3\wedge \text{train}(y)\wedge\text{tunnel}(z) \quad (5)$$

(S2) The path runs through the forest.

$$(\exists x,y,p_1,z,p_3,k,p_4,k_0)(\underline{L(x,y,p_1,z,A_{12},G_s,k)}\bullet L(x,y,z,p_3,A_{12},G_s,k))\Pi L(x,y,p_4,p_4,A_{13},G_s,k_0) \wedge p_1\neq z\wedge z\neq p_3\wedge \text{path}(y)\wedge\text{forest}(z) \quad (6)$$

(S3) The path through the forest is long.

$$(\exists x,y,p_1,z,p_3,x_1,k,q,k_1,p_4,k_0) (L(x,y,p_1,z,A_{12},G_s,k)\bullet L(x,y,z,p_3,A_{12},G_s,k)) \Pi L(x,y,p_4,p_4,A_{13},G_s,k_0)\wedge L(x_1,y,q,q,A_{02},G_t,k_1) \wedge p_1\neq z\wedge z\neq p_3\wedge q>k_1\wedge\text{path}(y)\wedge\text{forest}(z) \quad (7)$$

$$(\exists x_1,y_1,q,k_1)L(x_1,y_1,q,q,A_{02},G_t,k_1)\wedge q>k_1 \quad (8)$$

$$(\exists x_1,y_1,k_1)L(x_1,y_1,\text{Long},\text{Long},A_{02},G_t,k_1) \quad (8')$$

Every version of our intelligent system IMAGES [e.g.,4] can perform text understanding based on word meaning descriptions as follows.

Firstly, a text is parsed into a surface dependency structure (or more than one if *syntactically* ambiguous). Secondly, each surface dependency structure is translated into a conceptual structure (or more than one if *semantically* ambiguous) using word-meaning descriptions. Finally, each conceptual structure is semantically evaluated.

The fundamental semantic computations on a text are to detect semantic anomalies, ambiguities and paraphrase relations. Semantic anomaly detection is very important to cut off meaningless computations. Consider such a conceptual structure as (9), where 'A_{29}' is the attribute 'Taste'. This locus formula can correspond to the English sentence 'The desk is sweet', which is usually semantically anomalous because a 'desk' *ordinarily* has no taste.

$$(\exists x)L(\_,x,\text{Sweet},\text{Sweet},A_{29},G_t,\_)\wedge\text{desk}(x) \quad (9)$$

This kind of semantic anomaly can be detected in the following process. Firstly, assume the commonsense knowledge of 'desk' as (10), where 'A_{39}' refers to the attribute 'Vitality'. The special symbols '*' and '/' are defined as (11) and (12) representing 'always' and 'no value', respectively. The anonymous variable '_' defined by (13) is often used instead of the variable bound by an existential quantifier.

$$(\lambda x)\ \text{desk}(x)\leftrightarrow(\lambda x)(\ldots L^*(\_,x,/,/,A_{29},Gt,\_) \wedge\ldots\wedge L^*(\_,x,/,/,A_{39},Gt,\_)\wedge\ldots) \quad (10)$$

$$X^*\leftrightarrow(\forall t_1,t_2)X\,\Pi\,\varepsilon(t_1,t_2) \quad (11)$$

$$L(\ldots,/,\ldots)\leftrightarrow\sim(\exists p)\,L(\ldots,p,\ldots) \quad (12)$$

$$L(\ldots,\_,\ldots)\leftrightarrow(\exists x)L(\ldots,x,\ldots) \quad (13)$$

Secondly, the postulates (14) and (15) are utilized. The formula (14) means that ***if one of two loci exists every time interval, then they can coexist*** and the formula (15) states that ***a matter has never different values of an attribute at a time***.

$$X\wedge Y^*\ .\supset.\ X\ \Pi\ Y \quad (14)$$

$$L(x,y,p_1,q_1,a,g,k)\Pi L(z,y,p_2,q_2,a,g,k) .\supset.p_1=p_2\wedge q_1=q_2 \quad (15)$$

Lastly, the semantic anomaly of 'sweet desk' is detected by using (9)-(15). That is, the formula (16) below is finally deduced from (9)-(14) and violates the commonsense given by (15), that is, " *Sweet* ≠ / ".

$$(\exists x)L(\_,x,\text{Sweet},\text{Sweet},A_{29},G_t,\_)\Pi L(\_,x,/,/,A_{29},G_t,\_) \quad (16)$$

This process is also employed for dissolving such a syntactic ambiguity as found in S4. That is, the semantic anomaly of 'sweet desk' is detected and eventually 'sweet coffee' is adopted as a plausible interpretation.

(S4) Bring me the coffee on the desk, which is sweet.

If a text has multiple plausible interpretations, it is semantically ambiguous. In this case, IMAGES will ask for further information in order for disambiguation.

For another case, if two different texts are interpreted into the same locus formula, they are paraphrases of each other. The detection of paraphrase relations is very useful for deleting redundant information.

2. Pragmatic understanding

As mentioned above, an event expressed in L_{md} is compared to a movie film recorded through a floating camera because it is necessarily grounded in FAO's movement over the event ***This implies that L_{md} expression can suggest a robot what and how should be attended to in its environment.*** For example, consider such a suggestion as S5 presented to a robot by a human. In this case, unless the robot is aware of the existence of a certain box between the stool and the desk, such semantic understanding of the underlined part as (17) and such a semantic definition of the word 'box' as (18) are very helpful for it. The attributes A_{12} (Location), A_{13} (Direction), A_{32} (Color), A_{11} (Shape) and the spatial event on A_{12} in these L_{md} expressions indicate that the robot has only to activate its vision system in order to search for the box from the stool to the desk during the pragmatic understanding. That is, the robot can attempt to understand pragmatically the words of objects and events in an integrated top-down way.

(S5) Avoid <u>the green box between the stool and the desk</u>.

$$(\exists x_1,x_2,x_3,x_4,p)(L(\_,x_4,x_1,x_2,A_{12},G_s,\_)\bullet L((\_,x_4,x_2,x_3,A_{12},G_s,\_)) \Pi L(\_,x_4,p,p,A_{13},G_s,\_)\Pi L(\_,x_2,\text{Green},\text{Green},A_{32},G_t,\_) \wedge\text{stool}(x_1)\wedge\text{box}(x_2)\wedge\text{desk}(x_3)\wedge\text{ISR}(x_4) \quad (17)$$

$$(\lambda x)\text{box}(x)\leftrightarrow(\lambda x)L(\_,x,\text{Hexahedron},\text{Hexahedron},A_{11},G_t,\_) \wedge\text{container}(x) \quad (18)$$

III. BEHAVIOURALIZATION

The process for behaviouralization is to translate a conception (i.e., $\mathbf{C}_i$) into an action (i.e., $\mathbf{A}_i$) as an appropriate sequence of control codes for certain sensors or actuators in the robot to be decoded into a real behaviour. For this purpose, there are needed two kinds of core procedures so called 'Locus formula paraphrasing' and 'Behaviour chain alignment' as detailed below.

The attributes suggested by human are essentially for human sensors or actuators and therefore the locus formula as $\mathbf{C}_i$ should be translated into its equivalent concerning the attributes specific to the robot's. For example, an atomic locus of the robot's 'Shape (A_{11})' specified by the human should be paraphrased into a set of atomic loci of the 'Angularity (A_{45})' of each joint in the robot. For another example, 'Velocity (A_{16})' for the human into a set of change rates in 'Angularity (A_{45})' over 'Duration (A_{35})' (i.e., A_{45}/A_{35}) of the robot's joints involved. These knowledge pieces are called 'Attribute Paraphrasing Rules (APRs)' [4] and contained in $\mathbf{K}_D$.
Ideally, the atomic loci in the conception $\mathbf{C}_i$ (original or paraphrased) should be realized as the action $\mathbf{A}_i$ in a perfect correspondence with an appropriate chain of sensor or actuator deployments. Actually, however, such a chain as a direct translation of $\mathbf{C}_i$ must often be aligned to be feasible for the robot due to the situational, structural or functional differences between the human and the robot. For example of situational difference, consider the scene shown in Fig.6 where the robot Robby must interpolate the travel from its initial location to the green box and the action to pick up the box when the human Tom suggests him/it "Go to the table with the box between us." On the other hand, for example of structural or functional difference, consider the case of interaction between a human and a non-humanoid robot. Fig.7 shows the action by a dog-shaped robot (SONY) to the suggestion 'Walk and wave your left hand.' The robot pragmatically understood the suggestion as '*I* walk and wave *my* left *foreleg*' based on the knowledge piece that only forelegs can be waved' and behaviouralized its conception as 'I walk *BEFORE* sitting down *BEFORE* waving my left foreleg' but not as 'I walk, *SIMULTANEOUSLY* waving my left foreleg', in order not to fall down.

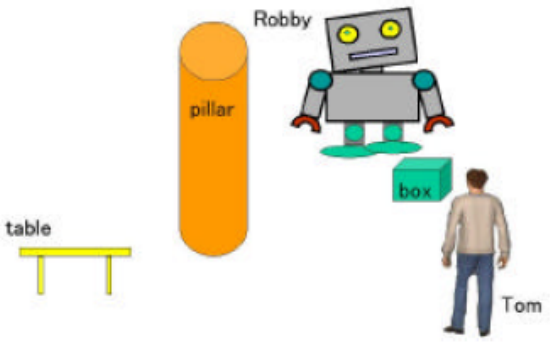

Fig.6. A scene of the robot Robby and the human Tom

Fig.7. Robot's action to the suggestion 'Walk and wave your left hand.'

IV. CONCLUSION

The key contribution of this paper is the proposal of a novel idea of robot manipualtion driven by semantic representation of human suggestion, where are hinted in the formal language $\boldsymbol{L}_{md}$ what and how should be attended to as analogy of human FAO movement and thereby the artificial attention can be controlled in a top-down way. The authors have a good perspective for the proposed theory of language-centered human-robot interaction based on their previous work utilizing $\boldsymbol{L}_{md}$ for robot manipulation by text [8]. This is one kind of cross-media operation by integrated multimedia understanding based on intermediate $\boldsymbol{L}_{md}$ representation [9]. At our best knowledge, there is no other theory or system that can perform cross-media operations in such a seamless way as ours. Our future work will include establishment of learning facilities for automatic acquisition of word concepts from sensory data and multimodal interaction between humans and robots under real environments in order to realize the robot manipulation proposed here.

REFERENCES

[1] Coradeschi S, Saffiotti A (2003), An introduction to the anchoring problem. Robotics and Autonomous Systems, 43: 85–96
[2] Drumwright E, Ng-Thow-Hing V, Mataric´ M J (2006), Toward a vocabulary of primitive task programs for humanoid robots. Proc. of International Conference on Development and Learning (ICDL), Bloomington, IN
[3] Yokota M (2005), An approach to integrated spatial language understanding based on Mental Image Directed Semantic Theory. Proc. of 5th Workshop on Language and Space, Bremen, Germany
[4] Yokota M, Capi G (2005), Cross-media Operations between Text and Picture Based on Mental Image Directed Semantic Theory. WSEAS Transactions on Information Science and Applications, 10-2: 1541-1550
[5] Yokota M (2006), Towards a Universal Knowledge Representation Language for Ubiquitous Intelligence Based on Mental Image Directed Semantic Theory. J.Ma et al.(Eds.) Ubiquitous Intelligence and Computing 2006 (UIC 2006), LNCS 4159: 1124-1133
[6] Sowa J F (2000), Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA
[7] Langacker R (1991), Concept, Image and Symbol. Mouton de Gruyter, Berlin/New York
[8] Yokota M (2006), Towards a universal language for distributed iIntelligent robot networking. Proc. of 2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan
[9] Yokota M (2008) Intuitive spatiotemporal representation based on Mental Image Description Language ***Lmd***. The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08), Beppu, Oita, Japan