Modeling of human mental-image based understanding of spatiotemporal language for intuitive human-robot interaction

Masao Yokota
Department of System Management, Fukuoka Institute of Technology, Fukuoka, Japan
yokota@fit.ac.jp

Rojanee Khummongkol
Department of Computer Engineering, University of Phayao, Phayao, Thailand
rojanee.kh@up.ac.th

Abstract

Mental Image Directed Semantic Theory (MIDST) has proposed a human mental image model and its description language $\mathbf{L_{md}}$. This is one kind of knowledge representation language and has already been applied to integrative multimedia understanding intended for facilitating intuitive human-robot interaction, especially, language-centered interaction between ordinary people and home robots. The most remarkable feature of $\mathbf{L_{md}}$ is its capability of formalizing spatiotemporal events in good correspondence with human/robotic sensations and actions, which can lead to integrative computation of sensory, motory and conceptual information. This paper sketches MIDST and its application, namely, the natural language understanding system named conversation management system (CMS) intended to simulate human mental-image based understanding of natural language, overviewing related work. CMS was evaluated based on a psychological experiment and showed a good agreement with human subjects in answering questions about stimulus sentences, inevitably involving spatiotemporal reasoning.

2012 ACM Subject Classification General and reference → General literature; General and reference

Keywords and phrases Natural language understanding, Mental image model, Human-robot interaction, Knowledge representation,Spatiotemporal reasoning.

Acknowledgements This work was partially funded by the Grants from Ministry of Education, Culture, Sports, Science and Technology, Japanese Government, numbered 14580436, 17500132, 23500195 and 19K12109.

1 Introduction

For ordinary people, natural language (NL) is the most important among the various communication media because it can convey the exact intention of the emitter to the receiver due to the syntax and semantics common to its users. This is not necessarily the case for another media, such as gesture, and so NL can also play the most crucial role in intuitive human-robot interaction (iHRI) intended here and shown in Figure 1. This figure implies that the robot should find and solve the problems in knowledge representation language (KRL) communicating with the human in NL. As easily understood, in such a scenario, the robot must be provided with a very powerful artificial intelligence (AI) for integrative comprehension of perceptual information (i.e., sensory or motory data) and conceptual information (i.e., lexical knowledge or world knowledge), and, especially, its capability of natural language understanding (NLU) (or more broadly, natural language processing (NLP)) should be much more cognitively elaborated than the conventional approaches (e.g., [26]; [19]; [7]; [27]) in order to cope with symbol grounding problems ([8]).

In the field of ontology, special attention has been paid to spatial (more exactly, spatiotemporal) language covering geography because its constituent concepts stand in highly

In Proceedings Speaking of Location 2019: Communicating about Space, Regensburg, Germany, September-2019.
Editors: K. Stock, C.B. Jones and T. Tenbrink (eds.);
Published at http://ceur-ws.org

complex relationships to underlying physical reality, accompanied with fundamental issues in terms of human cognition (for example, ambiguity, vagueness, temporality, identity, ...) appearing in varied subtle expressions ([9]). For facilitating iHRI, spatial language is also the most important of all sublanguages, especially, when both the entities must share knowledge of spatial arrangement of home utilities such as desk, table, etc.

As known well, people do not perceive the external world as it is, which naturally leads to human-specific cognition and conception of the external world. For example, as shown in Figure 2, people often perceive continuous forms among separately located objects so called spatial gestalts in the field of psychology and refer to them by such an expression as 'Nine disks are placed in the shape of X'. For another example, people would intuitively and easily understand the following expressions S1 and S2 so that they describe the same scene in the external world. This is also the case for S3 and S4.

(S1) The path sinks to the brook.
(S2) The path rises from the brook.
(S3) The roads meet there.
(S4) The roads separate there.

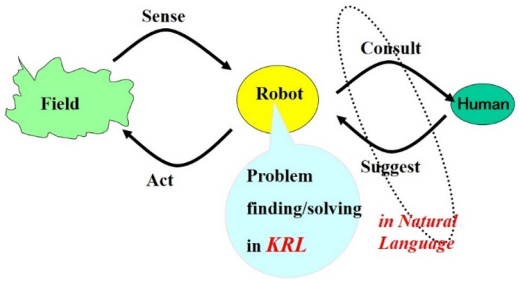

Figure 1 Intuitive human-robot interaction

Figure 2 Gestalt perceived among multiple disks

It is, however, extremely difficult for robots to reach such a paradoxical understanding in a systematic way because these expressions are assumed to reflect not so much the purely objective geometrical relations but very much human mental activity at cognition of the involved objects, inevitably employing mental image operations (e.g., [28], [30], [29], [32]). However, most conventional approaches to spatial language understanding have focused on computing ostensible geometric relations (i.e., topological, directional and metric relations) conceptualized as spatial prepositions or so, considering some properties or functions of the objects involved (e.g., [5]; [16]; [2]). From the semantic viewpoint, spatial expressions have the virtue of relating in some way to visual scenes being described. Therefore, their semantic descriptions can be grounded in perceptual representations, possibly, cognitively inspired and coping with all kinds of spatial expressions including such verb-centered ones as S1-S4 as well as preposition-centered ones. In particular, these verb-centered expressions are

assumed to reflect very much certain dynamism at human perception of the objects involved. This implies that conventional approaches to spatial language understanding will inevitably lead to serious cognitive divide between humans and robots that causes miscommunication between them. That is, AI should be more cognitive ([28]; [24]).

Reflecting our own psychological expriences, mental image must deeply concern our thinking. It is considered that people can create fictive or non-veridical stories thanks to mental images independent of the real world. More actually, it is quite ordinary to understand a spatiotemporal (or 4D) expression in NL with the mental image of a certain scene being described by it. Therefore, such a human mental process is worth simulating by computers in order to facilitate iHRI.

MIDST (Mental Image Directed Semantic Theory) (e.g., [28]; [32]) has proposed a dynamic model of human sensory cognition yielding omnisensory image of the world. In MIDST, natural event concepts (i.e., event concepts in NL) are classified into two types of categories, 'Temporal Change Events' and 'Spatial Change Events'. These are defined as temporal and spatial changes (or constancies) in certain attributes of physical objects, respectively, with S1-S4 included in the latter. Both the types of events are uniformly analyzable as temporally parameterized loci in attribute spaces to be described distinctively in a logical form, so called, "locus formula". MIDST has already been applied to several types of computerized intelligent systems (e.g., [28]; [11]) and there is a feedback loop between them for their mutual refinement.

This paper sketches MIDST and our NLU system named conversation management system (CMS) intended to simulate human mental-image based understanding of natural language with its evaluation based on a psychological experiment. The remainder of this paper is organized as follows. Section II considers human mental image-based understanding of natural language and Section III presents a brief description of MIDST. Section IV describes the methodology for NLU based on MIDST. Section V gives a brief description of CMS and its evaluation based on a psychological experiment. Lastly, Section VI concludes this paper.

2 Mental-image based NLU

For example, read the assertion S5 and answer to the questions S6 and S7. Perhaps, without any exception, we cannot answer the questions correctly without reasoning based on the mental images evoked by these expressions.

(S5) Mary was in the tram heading for the town. She had a bag with her.

(S6) Was the tram carrying Mary?

(S7) Was the bag heading for the town?

This kind of reasoning is considered to belong to what are required for the Winograd Schema Challenge (WSC) ([15]) that would discourage conventional NLU systems adapted for the Turing Test, even the renowned quiz champion AI, Watson by IBM ([6]). The WSC is a more cognitively-inspired variant of the Textual Entailment Challenge ([3]) to eliminate cheap tricks intended to pass the Turing Test which is essentially based on behaviorism in the field of psychology.

There are a considerable number of cognitively motivated studies on NL semantics or pragmatics in association with mental image explicitly or implicitly (e.g., [14]; [17]; [21]; [22]; [13]). However, almost none of them are for NLU because certain systematic methodologies for both representation and computation of mental imagery are inevitably required for NLU. As well, a lot of interesting researches on mental image itself in association with human thinking modes have been reported from various fields ([23]) but none of them are from the

viewpoint of NLU, either.

Distinctively from them, MIDST is intended for systematic representation and computation of NL semantics, more broadly, human knowledge grounded in the world through mental images.

Originally, our mental images of the external world are acquired through our inborn sensory systems, and therefore, it is worth considering our perceptual processes. As already mentioned, we do not perceive the external world as it is. That is, our perception does not begin with objective data gained through artificial sensors but with subjective sensation intrinsically (or subconsciously) articulated with contours of involved objects and gestalts among them as shown in Figure 2. Here, this kind of articulation is called intrinsic articulation, attributed to our subconscious propensities toward the external world. Then, at the next stage, as active perception, we work our attention consciously to elaborate (or calibrate) intrinsic articulation by reasoning based on various kinds of knowledge and come to an interpretation of the sensation as a spatiotemporal relation among its significant portions or constituents. Here, this elaboration of intrinsic articulation is called semantic articulation that MIDST concerns in particular. The neural network architectures prevailing today are based on simple-minded algorithms and are essentially to provide a machine with intrinsic articulation but not semantic articulation of the stimuli posed.

Overviewing conventional methodologies for robotic NLU, almost all of them have provided robotic systems with such quasi-natural language expressions as 'move(Velocity, Distance, Direction)', 'find(Object, Shape, Color)', etc., for human instruction or suggestion, uniquely related to computer programs for deploying sensors/motors as their semantics (e.g., [1]; [4]). These expression schemas, however, are too linguistic or coarse to represent and compute sensory/motory events in such an integrative way as the intuitive human-robot interaction intended here. This is also the case for AI planning ('action planning'), which deals with the development of representation languages (i.e., KRLs) for planning problems and with the development of algorithms for plan construction ([25]).

In order to challenge a complex problem domain, the first thing to do is to design/select a certain KRL suitable for constructing a well-structured problem formulation, namely, a representation. Among conventional KRLs, the ones employable for first order logic have been the most prevailing because of good availability of deductive inference engines intrinsically prepared for computer languages such as Python. According to these schemes, for example, the semantic relation between 'x carry y' and 'y move' is often to be translated into such a representation as $(\forall x, y)(carry(x, y) \supset move(y))$. As easily imagined, such a declarative definition will enable an NLU system to answer correctly to such a question as "When Jim carried the box, did it move?" but it will be of no use for a robot to recognize or produce any external event referred to by 'x carry y' or 'y move' in a dynamic and incompletely known environment unlike the Winograd's block world ([26]). That is, this type of logical expression as is can give only combinations of dummy tokens at best. For example, $carry(x, y)$ and $move(y)$ are substitutable with $w013(x, y)$ and $w025(y)$, respectively, which do not represent any word concepts or meanings at all but are the coded names of such concepts or meanings. If you find any inconvenience with this kind of substitution, that is due to being without symbol grounding ([8]) on your lexical knowledge of English. Schank's Conceptual Dependency theory ([20]) was an attempt to decrease paraphrastic variety in knowledge representation by employing a small set of coded names of concepts called conceptual primitives, although its expressive power was very limited.

The fact above destines a cognitive robot to be provided with procedural definitions of word meanings grounded in the external world, as well as declarative ones for reasoning in

order both to work its sensors and actuators appropriately and to communicate by natural language with humans properly. Therefore, it is noticeable that some certain interface must be employed for translation between declarative and procedural definitions of word meanings, where the problem is how to realize such a translator systematically. Conventional KRLs, however, are not so viable of such systematization because they are not so cognitively designed, namely, not so systematically grounded in sensors or actors. That is, they are not provided with their semantics explicitly but implicitly grounded in natural language word concepts that can be interpretable for people but have never been grounded in the world well enough for robots to cognize their environments or themselves through NL expressions.

3 Brief Description of MIDST

MIDST has proposed a dynamic model of human sensory cognition, yielding omnisensory image of the world and its description language named 'mental image description language' ($\mathbf{L_{md}}$). This formal language is one kind of KRL employed for predicate logic. In MIDST, omnisensory mental images are modeled as "Loci in Attribute Spaces". An attribute space corresponds with a certain measuring instrument just like a barometer, thermometer or so and the loci represent the movements of its indicator.

For example, the moving red triangular object shown in Figure 3 is assumed to be perceived as the loci in the three attribute spaces, namely, those of 'Location', 'Color' and 'Shape' in the observer's brain as the result of intrinsic articulation. At the next stage as semantic articulation, a general locus is to be articulated by "Atomic Locus" as depicted in Figure 4 and formulated as (1).

$$L(x, y, p, q, a, g, k) \tag{1}$$

The intuitive interpretation of (1) is given as follows. **"Matter 'x' causes Attribute 'a' of Matter 'y' to keep (p=q) or change (p ≠ q) its values temporally ($g = G_t$) or spatially ($g = G_s$) over a time-interval, where the values 'p' and 'q' are relative to the standard 'k'."**

When $g = G_t$, the locus indicates monotonic change or constancy of the attribute in time domain and when $g = G_s$, that in space domain, respectively. The former is called 'temporal change event' and the latter, 'spatial change event'. For example, the motion of the 'bus' represented by S8 is a temporal change event and the ranging or extension of the 'road' by S9 is a spatial change event whose meanings or concepts are formulated as (2) and (3), respectively, where 'A12' denotes the attribute 'Physical Location'. These two formulas are different only at the term 'Event Type ($= g$)'.

(S8) The bus runs from Tokyo to Osaka.

$$(\exists x, y, k)L(x, y, Tokyo, Osaka, A_{12}, G_t, k) \wedge bus(y) \tag{2}$$

(S9) The road runs from Tokyo to Osaka.

$$(\exists x, y, k)L(x, y, Tokyo, Osaka, A_{12}, G_s, k) \wedge road(y) \tag{3}$$

The formal language Lmd has employed 'tempo-logical connectives' representing both logical and temporal relations between loci. Articulated loci are combined with tempo-logical conjunctions, where 'SAND ($\wedge_0$)' and 'CAND ($\wedge_1$)' are most frequently utilized, standing for 'Simultaneous AND' and 'Consecutive AND', conventionally symbolized as 'Π' and '·', respectively. For example, the expression (4) is the definition of the English verb concept

'fetch' depicted as Figure 5. This implies such a temporal change event that x goes for y and then comes back with it, where the special symbol λ is employed to denote free vaiables explicitly.

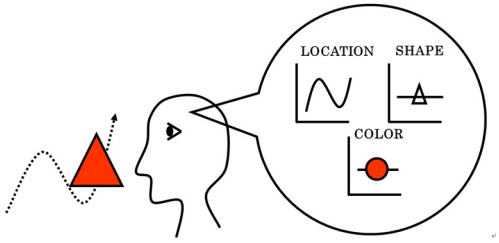

Figure 3 Mental image model as loci in attribute spaces

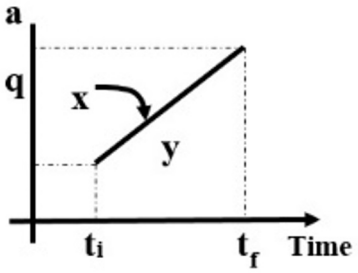

Figure 4 Atomic Locus in Attribute Space

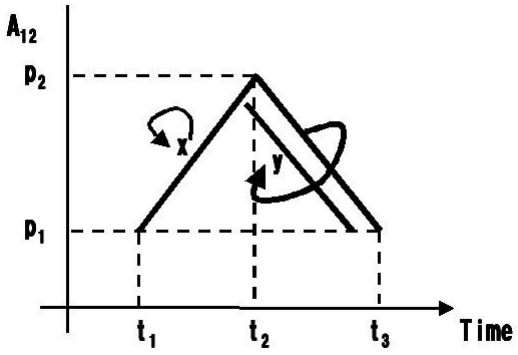

Figure 5 Image of the English verb 'fetch'

$$
\begin{aligned}
&(\lambda x, y) fetch(x, y) \\
&\Leftrightarrow (\lambda x, y)(\exists p_1, p_2, k) L(x, x, p_1, p_2, A_{12}, G_t, k) \cdot \\
&\qquad ((L(x, x, p_2, p_1, A_{12}, G_t, k) \Pi L(x, y, p_2, p_1, A_{12}, G_t, k)) \wedge x \neq y \wedge p_1 \neq p_2 \quad (4)
\end{aligned}
$$

It has been often argued that human active sensing processes may affect perception and in turn conceptualization and recognition of the physical world. The difference between temporal and spatial change event concepts can be attributed to the relationship between the Attribute Carrier (AC) and the Focus of Attention of the Observer (FAO). To be brief, the

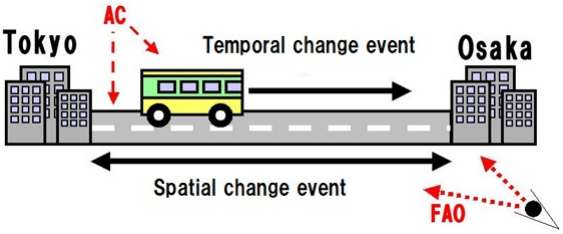

Figure 6 Event types and FAO movements

FAO is fixed at one point on the AC in a temporal change event but runs about on the AC in a spatial change event. Consequently, as shown in Figure 6, the bus and the FAO move together in the case of S8 while the FAO solely moves along the road in the case of S9. That is, **all loci in attribute spaces correspond one to one with movements or, more generally, temporal change events of the FAO. This implies that L_{md} expression can suggest a robot what and how should be attended to in its environment**. And this is why S1 and S2 (as well as S3 and S4) can refer to the same scene in spite of their appearances, where what 'sinks' or 'rises' is the FAO and whose conceptual descriptions are given as (5) and (6), respectively, where 'A13', '↑' and '↓' refer to the attribute 'Direction' and its values 'upward' and 'downward', respectively. Such a fact is generalized as '**Postulate of Reversibility of a Spatial Change Event** (PRS)' that can be one of the principal cognitive laws belonging to people's intuitive common-sense knowledge about geography. These pairs of conceptual descriptions are called **equivalent in the PRS**, and the paired sentences are treated as **paraphrases** each other.

$$(\exists x, y, p, z, k_1, k_2)L(x, y, p, z, A_{12}, G_s, k_1)\Pi \\ L(x, y, \downarrow, \downarrow, A_{13}, G_s, k_2) \wedge path(y) \wedge brook(z) \wedge p \neq z \quad (5)$$

$$(\exists x, y, p, z, k_1, k_2)L(x, y, z, p, A_{12}, G_s, k_1)\Pi \\ L(x, y, \uparrow, \uparrow, A_{13}, G_s, k_2) \wedge path(y) \wedge brook(z) \wedge p \neq z \quad (6)$$

For another example of spatial change event, Figure 7 concerns the perception of the formation of multiple isolated objects, where FAO runs along an imaginary object so called 'Imaginary Space Region (ISR)'. This spatial event can be verbalized as S10 using the preposition 'between' and formulated as (7) or (8), corresponding also to such concepts as 'row', 'line-up', etc. It is noticeable that ISR is itended to include spatial gestalt conceptually but it is assumed that people can imagine ISRs consciously or arbitrarily at semantic articulation..

(S10) Y is between X and Z.

$$(\exists x, y, p, q, k_1, k_2)(L(x, y, X, Y, A_{12}, G_s, k_1)\Pi L(x, y, p, p, A_{13}, G_s, k_2)) \\ \cdot (L(x, y, Y, Z, A_{12}, G_s, k_1)\Pi L(x, y, q, q, A_{13}, G_s, k_2)) \wedge ISR(y) \wedge p = q \quad (7)$$

$$(\exists x, y, p, k_1, k_2)(L(x, y, Z, Y, A_{12}, G_s, k_1)\cdot \\ L(x, y, Y, X, A_{12}, G_s, k_1))\Pi L(x, y, p, p, A_{13}, G_s, k_2) \wedge ISR(y) \quad (8)$$

At our best knowledge, there is no other theory or method that can provide spatiotemporal expressions with semantic interpretation in such a systematic way where both temporal and

spatial change events are simply and adequately formulated by controlling the term of Event Type of the atomic locus formula reflecting FAO movement. About 50 attributes (Table 1) were extracted exclusively from English and Japanese words of common use contained in certain thesauri (e.g., [18]). Most of them correspond to the sensory receptive fields in human brains. Correspondingly, seven categories of standards (Table 2) were extracted that are necessary for representing relative values of each attribute.

These findings imply that ordinary people live their casual life, attending to tens of attributes of the matters in the world to cognize them in comparison with several kinds of standards. That is, with verbal hint, a robot can work its sensors or actuators very efficiently or economically and otherwise it is extremely difficult for the robot to understand which part of its environment is significant or not for people because there are too many things to attend to as is.

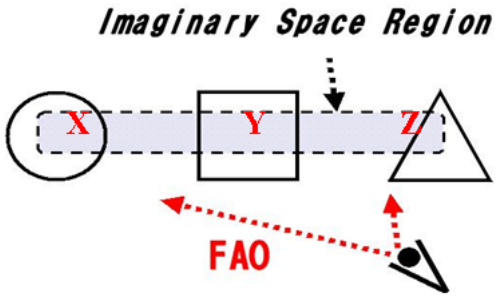

Figure 7 Spatial event 'row' and FAO movement

4 NLU Based on $\mathbf{L_{md}}$

In MIDST, natural language expression (i.e., surface structure) and Lmd expression (i.e., conceptual structure) are mutually translatable through the surface dependency structure by utilizing syntactic rules and word meaning descriptions (e.g., [28]).

A word meaning description Mw is given by (9) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$$M_w \Leftrightarrow [C_p : U_p] \tag{9}$$

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 'x carry y' is approximately given by (10), where A_{12} is the attribute of "Physical location".

$$\begin{aligned}&[(\lambda x, y)(\exists p_1, p_2, k)L(x, x, p_1, p_2, A_{12}, G_t, k)\Pi\\ &\quad L(x, y, p_1, p_2, A_{12}, G_t, k) \wedge x \neq y \wedge p_1 \neq p_2 : ARG(Dep.1, x); ARG(Dep.2, y);]\end{aligned} \tag{10}$$

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 to be translated into (11) via

Table 1 List of attributes

Cou	Attribute	[Property]	(words/phrases)
$*A_{01}$	WORLD	[N]	(in the dream, imaginary)
$*A_{02}$	LENGTH	[S]	(long, shorten, close, away)
$*A_{03}$	HEIGHT	[S]	(high, lower)
$*A_{04}$	WIDTH	[S]	(widen, narrow)
$*A_{05}$	THICKNESS	[S]	(thick, thin)
$*A_{06}$	DEPTH1	[S]	(deep, shallow)
$*A_{07}$	DEPTH2	[S]	(deep, concave)
$*A_{08}$	DIAMETER	[S]	(across, in diameter)
$*A_{09}$	AREA	[S]	(square meters, acre)
$*A_{10}$	VOLUME	[S]	(litter, gallon)
$*A_{11}$	SHAPE	[N]	(round, triangle)
$*A_{12}$	PHYSICAL LOCATION	[N]	(move, stay)
$*A_{13}$	DIRECTION	[N]	(turn, wind, left)
$*A_{14}$	ORIENTATION	[N]	(orientate, command)
$*A_{15}$	TRAJECTORY	[N]	(zigzag, circle)
$*A_{16}$	VELOCITY	[S]	(fast, slow)
$*A_{17}$	MILEAGE	[S]	(far, near)
A_{18}	STRENGTH OF EFFECT	[S]	(strong, powerful)
A_{19}	DIRECTION OF EFFECT	[N]	(pull, push)
A_{20}	DENSITY	[S]	(dense, thin)
A_{21}	HARDNESS	[S]	(hard, soft)
A_{22}	ELASTICITY	[S]	(elastic, flexible)
A_{23}	TOUGHNESS	[S]	(fragile, stiff)
A_{24}	TACTILE FEELING	[S]	(rough, smooth)
A_{25}	HUMIDITY	[S]	(wet, dry)
A_{26}	VISCOSITY	[S]	(oily, watery)
A_{27}	WEIGHT	[S]	(heavy, light)
A_{28}	TEMPERATURE	[S]	(hot, cold)
A_{29}	TASTE	[N]	(sour, sweet, bitter)
A_{30}	ODOUR	[N]	(pungent, sweet)
A_{31}	SOUND	[N]	(noisy, silent, loud)
$*A_{32}$	COLOR	[N]	(red, white)
A_{33}	INTERNAL SENSATION	[N]	(tired, hungry)
A_{34}	TIME POINT	[S]	(o'clock, elapse)
A_{35}	DURATION	[S]	(hour, minute, long, short)
A_{36}	NUMBER	[S]	(ten, quantity, number)
A_{37}	ORDER	[S]	(first, last)
A_{38}	FREQUENCY	[S]	(sometimes, frequent)
A_{39}	VITALITY	[S]	(alive, dead, vivid)
A_{40}	SEX	[S]	(male, female)
A_{41}	QUALITY	[N]	(make, destroy)
A_{42}	NAME	[V]	(name, token)
A_{43}	CONCEPTUAL CATEGORY	[V]	(mammal)
$*A_{44}$	TOPOLOGY	[V]	(in, out, touch)
$*A_{45}$	ANGULARITY	[S]	(sharp, dull, rectangle)
B_{01}	WORTH	[N]	(improve, praise, deny, alright)
B_{02}	LOCATION OF INFORMATION	[N]	(tell, hear)
B_{03}	EMOTION	[N]	(like, hate)
B_{04}	BELIEF VALUE	[S]	(believe, trust)

Table 2 List of standards

Categories	Remarks
Rigid Standard	Objective standards such as denoted by measuring units (meter, gram, etc.).
Species Standard	The attribute value ordinary for a species. A short train is ordinarily longer than a long pencil.
Proportional Standard	'Oblong' means that the width is greater than the height at a physical object.
Individual Standard	Much money for one person can be too little for another.
Purposive Standard	One room large enough for a person's sleeping must be too small for his jogging.
Declarative Standard	The origin of an order such as 'next' must be declared explicitly just as 'next to him'.
Tacit Standard	Gives granularities for semantic articulation of loci. Most of them are tacit due to non-linguistic cognitive processes working as the units of cognitive scales.

dependency structure and vice versa as depicted in Figure 8.

$$(\exists y, p_1, p_2, k)L(Mary, Mary, p_1, p_2, A_{12}, G_t, k)\Pi L(Mary, y, p_1, p_2, A_{12}, G_t, k) \\ \wedge Mary \neq y \wedge p_1 \neq p_2 \wedge book(y) \quad (11)$$

For another example, the meaning description of the English preposition 'x (verb)

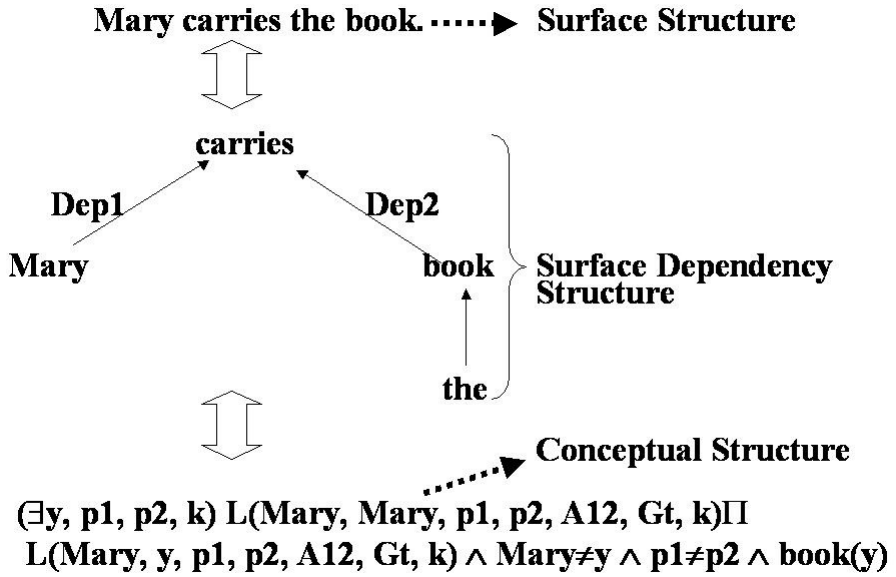

Figure 8 Mutual conversion between NL text and $\mathbf{L_{md}}$

through y' is also approximately given by (12).

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

The U_p above is for unifying the $C_p s$ 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 (12) 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 $C_p s$ to be unified and called 'Unification Handle (U_h)' and when missing, the $C_p s$ are to be combined simply with '$\wedge$'.

Therefore the sentences S11-S13 are interpreted as (13)-(15), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (16) 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 (17).

(S11) The train runs through the tunnel.

$$(\exists x, y, p_1, z, p_3, k, p_4, k_0)(\underline{L(x, y, p_1, z, A_{12}, G_t, k)} \cdot L(x, y, z, p_3, A_{12}, G_t, k))$$
$$\Pi L(x, y, p_4, p_4, A_{13}, G_t, k_0) \wedge p_1 \neq z \wedge z \neq p3 \wedge train(y) \wedge tunnel(z) \quad (13)$$

(S12) The path runs through the forest.

$$(\exists x, y, p_1, z, p_3, k, p_4, k_0)(\underline{L(x, y, p_1, z, A_{12}, G_s, k)} \cdot L(x, y, z, p_3, A_{12}, G_s, k))$$
$$\Pi L(x, y, p_4, p_4, A_{13}, G_s, k_0) \wedge p_1 \neq z \wedge z \neq p3 \wedge path(y) \wedge forest(z) \quad (14)$$

(S13) The path through the forest is long.

$$(\exists x, y, p_1, z, p_3, x_1, k, q, k_1, p_4, k_0)(L(x, y, p_1, z, A_{12}, G_s, k)\cdot$$
$$L(x, y, z, p_3, A_{12}, G_s, k))\Pi L(x, y, p_4, p_4, A_{13}, G_s, k_0) \wedge L(x_1, y, q, q, A_{02}, G_t, k_1)$$
$$\wedge p1 \neq z \wedge z \neq p_3 \wedge q > k_1 \wedge path(y) \wedge forest(z) \quad (15)$$

$$(\exists x_1, y_1, q, k_1)L(x1, y_1, q, q, A_{02}, G_t, k_1) \wedge q > k_1 \quad (16)$$

$$(\exists x_1, y_1, k_1)L(x1, y1, Long, Long, A_{02}, G_t, k_1) \quad (17)$$

Every version of our intelligent system IMAGES (e.g., [28]; [11]) 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 ([10]). Semantic anomaly detection is very important to prevent robots from meaningless computations and actions to such a verbal command by people as S14.

(S14) Find a moving object which is stationary.

Consider such a conceptual structure as (18), where 'A_{29}' is the attribute 'Taste' and 'Sw' is the value for 'sweetness'. 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. The anonymous variable '_' defined by (22) is often used instead of the variable bound by an existential quantifier, for the sake of simplicity.

$$(\exists x)L(\_, x, Sw, Sw, A_{29}, G_t, \_) \wedge desk(x) \tag{18}$$

This kind of semantic anomaly can be detected in the following process. Firstly, assume the commonsense knowledge of 'desk' as (16), where 'A_{39}' refers to the attribute 'Vitality'. The special symbols '*' and '/' are defined as (20) and (21) representing 'always' and 'no value', respectively.

$$(\exists x)desk(x) \leftrightarrow (\lambda x)(\dots L^*(\_, x, /, /, A_{29}, G_t, \_) \wedge \dots \wedge L^*(\_, x, /, /, A_{39}, G_t, \_) \wedge \dots) \tag{19}$$

$$X^* \leftrightarrow (\forall t_1, t_2)X\Pi\varepsilon(t_1, t_2) \tag{20}$$

$$L(\dots, /, \dots) \leftrightarrow\sim (\exists p)L(\dots, p, \dots) \tag{21}$$

$$L(\dots, \_, \dots) \leftrightarrow (\exists x)L(\dots, x, \dots) \tag{22}$$

Secondly, the postulates (23) and (24) are utilized. The formula (23) means that *if one of two loci exists every time interval, then they can coexist* and the formula (24) states that *a matter never has different values of an attribute at a time.*

$$X \wedge Y^*. \supset .X\Pi Y \tag{23}$$

$$L(x, y, p_1, q_1, a, g, k)\Pi L(z, y, p_2, q_2, a, g, k). \supset .p1 = p2 \wedge q1 = q2 \tag{24}$$

Lastly, the semantic anomaly of 'sweet desk' is detected by using (18)-(24). That is, the formula (25) below is finally deduced from (18)-(23) and violates the commonsense given by (24), that is, "$Sw \neq /$".

$$(\exists x)L(\_, x, Sw, Sw, A_{29}, G_t, \_)\Pi L(\_, x, /, /, A_{29}, G_t, \_) \tag{25}$$

This process is also employed for dissolving such a syntactic ambiguity as found in S15. That is, the semantic anomaly of 'sweet desk' is detected and eventually 'sweet coffee' is adopted as a plausible interpretation.

(S15) 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.

5 Conversation Management System

Our conversation management system (CMS), the latest version of IMAGES, understands User's assertions or questions in $\mathbf{L_{md}}$ and responds to them by text or animation. The general performances of CMS have already been published in ([11]) and, therefore, here is focused on how well it can simulate human mental-image based understanding (MBU) of spatiotemporal expresssions in NL. This capability was evaluated based on a psychological experiment and showed a good agreement with human subjects in answering questions about stimulus sentences, inevitably involving spatiotemporal reasoning.

The stimulus sentences to CMS and human subjects were I1-I3 as shown below.

(I1) Tom was with the book in the bus running from Town to University.

(I2) Tom was with the book in the car driven from Town to University by Mary.

(I3) Tom kept the book in a box before he drove the car from Town to University with the box.

Table 3 shows the questions about I1 - I3 and the answers by CMS which agreed with those of the human subjects very well. The human subjects were native speakers of Japanese, English, Chinese, or Thai. They were asked to sketch their mental images evoked by the stimulus sentences and to answer questions about them. As well, some disagreements were detected among all the participants (including CMS) and were found due to slight differences in their own word concepts. For example, one human subject answered only 'Tom' suitable for Q3 of I1 because his concept of the verb travel was limitedly applicable to human agents who move from one place to another (for a special purpose such as sightseeing).

In order to answer these questions, CMS performed reasoning based on several postulates as part of people's commonsense knowledge about the 4D world. They are P_{MV} (Postulate of Matter as Value), P_{SC} (Postulate of Shortcut in Causal Chain) and P_{CV} (Postulate of Conservation of Values in Time) as follows, where $\Lambda_t = (A_{12}, G_t, k)$.

P_{MV}	$(\forall \dots) L(z,x,p,q,\Lambda_t)\Pi L(w,y,x,x,\Lambda_t) \rightarrow L(z,x,p,q,\Lambda_t)\Pi L(w,y,p,q,\Lambda_t)$
P_{SC}	$(\forall \dots) L(z,x,p,q,\Lambda_t)\Pi L(w,y,x,x,\Lambda_t) \rightarrow L(z,x,p,q,\Lambda_t)\Pi L(z,y,p,q,\Lambda_t)$
P_{CV}	$(\forall \dots) L(z,x,p,p,\Lambda_t) \cdot X \rightarrow L(z,x,p,p,\Lambda_t) \cdot (L(z,x,p,p,\Lambda_t)\Pi X)$

For example, when $p \neq q$, P_{MV} reads that if 'z causes x to move from p to q as w causes y to be with x' then 'w causes y to move from p to q'. Similarly, P_{SC}, so that if 'z causes x to move from p to q as w causes y to be with x' then 'z causes y to move from p to q as well as x'. Distinctively from these two, P_{CV} is conditional, reading that if 'z keeps x at p (until some event X happens)' then 'it will continue'. That is, P_{CV} is valid only when X does not contradict with $L(z,x,p,p,\Lambda_t)$. These postulates are also applicable to the scene being described by S5 to answer S6 and S7.

6 Discussion

Katz and Fodor ([10]) presented the first analytical issue with human semantic processing ability. They claimed after their own experiences that people (more specifically, fluent speakers) can at least detect in a text or between texts such semantic properties or relations as follows and presented a model of disambiguation process called 'selection restriction' employing lexical information roughly specified by semantic markers and distinguishers in English.

(a) semantic ambiguity

(b) semantic anomaly
(c) paraphrase relation (i.e., semantic identity between different expressions)

To our best knowledge, there has been no systematic implementation of these functions reported in any NLU or NLP systems other than the work by us ([28]; [11]). Among them, the most essential for NLU is to detect paraphrase relation because the other two are possible if it is possible to determine equality (or inequality) between knowledge representations (or semantic representations) of different NL expressions.

As easily understood, the quality of this function depends on the capability of the adopted KRL to normalize knowledge representations, that is, to assign one knowledge representation to the same meanings. However, reflecting our psychological experiences in NLU, we utilize tacit or explicit knowledge associated to the words or so involved in order to process an NL expression semantically. This is also the case for NLU systems. That is, they should be inevitably provided with knowledge good enough for the purpose, for example, lexical and ontological knowledge, computably formalized in KRL and $\mathbf{L_{md}}$ can be a KRL appropriate enough for such a purpose as shown by the result of our psychological experiment.

The system CMS was designed to disambiguate an input sentence for its most plausible semantic interpretation by semantic computation (i.e., inference) in $\mathbf{L_{md}}$. Our psychological experiment revealed that the human subjects remembered their own experiences in association with the entity names and that they selected the dependency corresponding to their most familiar experience among all the possibilities. For example, consider the stimulus sentence I1. How can the machine know who/what was running from Town to University? —Tom, or book, or bus? Here, to see its syntactic possibilities, Dependency Grammar is employed in order to determine the relations between head words and their dependents. In principle, I1 can have twelve possible dependency trees, that is, it can be syntactically ambiguous in twelve ways. In this case, the names (e.g., Tom, bus, book) made the people remember the images in the way as formulated by (26) – (28), where $A \approx> B$ reads that A evokes B, and + and - denote whether the image is positive (i.e., probable) or negative (i.e., improbable), respectively.

$$Tom \approx> \{+L(\_, Tom, Human, Human, \Theta_t), +L(Tom, Tom, p, q, \Lambda_t), \dots\} \tag{26}$$

$$Book \approx> \{-L(Book, Book, p, q, \Lambda_t), +L(Human, Book, Human, Human, \Lambda_t), \dots\} \tag{27}$$

$$Bus \approx> \{+L(Bus, Bus, p, q, \Lambda_t), +L(Bus, x, p, q, \Lambda_t), +L(\_, Human, Bus, Bus, \Lambda_t), \dots\} \tag{28}$$

In (26), Θ_t represents 'Quality (or Category) (i.e., A_{41})' with $g = G_t$, and then $+L(\_, Tom, Human, Human, \Theta_t)$ is interpretable as 'it is positive that Tom is a human'. In the same way, $+L(Tom, Tom, p, q, \Lambda_t)$ as 'it is positive that Tom moves by himself', and $-L(Book, Book, p, q, \Lambda_t)$ as 'it is negative that a book moves by itself'. According to semantic preferences such as (26)-(28), CMS infers that the book did not run but Tom or bus and it reaches the final decision that the bus did because Tom was static in the bus.

Disambiguation is the most serious problem for any NLP system. Most current approaches to it are based on the statistics about certain corpora of texts, however, they are what lead to the most plausible syntactic interpretation but not to the most plausible semantic interpretation grounded in the concerned world that is most essential to work robots appropriately by

words. Concerning the research field of spatial language understanding, for example, the task of spatial role labeling ([12]) is intended to formalize the representation of spatial concepts and relations in the natural language text to be mapped to qualitative spatial representation models by means of machine learning techniques. This is in the same line as the UIMA (Unstructured Information Management Architecture) approach employed for Watson ([7]), specialized to extract spatial information from natural language texts, but its applicability to disambiguation or deeper understanding like ours remains questionable because it is to return spatial representations approximated by coded names at best.

7 Conclusion

Robotic NLU intended to simulate human NLU based on mental images was described. CMS at the present stage has been evaluated in comparison with human subjects for our psychological experiment on MBU and has shown a good agreement with them in NLU performance. The semantic computation in $\mathbf{L_{md}}$ performed by CMS is based on simple and general rules about atomic loci, hence, CMS works feasibly in Python except for computational cost in the Animation Generator. As for the coverage of word concepts, a considerable number of spatial terms have been analyzed over various kinds of English words, such as prepositions, verbs, adverbs, etc., categorized as Dimensions, Form and Motion in the class SPACE of the Roget's thesaurus, and it is found that almost all the concepts of 4D events can be defined in exclusive use of 5 kinds of attributes for FAO (the focus of attention of the observer), namely, Physical location (A_{12}), Direction (A_{13}), Trajectory (A_{15}), Mileage (A_{17}) and Topology (A_{44}). This implies that spatiotemporal information systems with NL interfaces are very feasible in terms of the size of knowledge to be installed.

The future work will include development of learning facilities for automatic acquisition of word concepts from sensory data and through language-centered interaction between humans and robots under real environments in the same way as human acquisition of language. On the other hand, the semantic plausibilities of names denoted as (26)-(28) can be more efficiently and automatically obtained from certain corpora of big size and good quality.

References

1 Silvia Coradeschi and Alessandro Saffiotti. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2):85 – 96, 2003. Perceptual Anchoring: Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems. URL: `http://www.sciencedirect.com/science/article/pii/S0921889003000216`, `doi:https://doi.org/10.1016/S0921-8890(03)00021-6`.

2 Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376 – 398, 2001. URL: `http://www.sciencedirect.com/science/article/pii/S0749596X00927426`, `doi:https://doi.org/10.1006/jmla.2000.2742`.

3 Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

4 Evan Drumwright, Victor Ng-Thow-Hing, and Maja J. Mataric. Toward a vocabulary of primitive task programs for humanoid robots. 2006.

5 Max J. Egenhofer and Robert D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991. URL: `https://doi.org/`

10.1080/02693799108927841, arXiv:https://doi.org/10.1080/02693799108927841, doi:10.1080/02693799108927841.

6 D Ferrucci, E Brown, J Chu-Carroll, J Fan, D Gondek, A A Kalyanpur, A Lally, J W Murdock, E Nyberg, J Prager, N Schlaefer, and C Welty. Building Watson: An Overview of the DeepQA Project. *AI MAGAZINE*, 31:59–79, 2010. URL: https://doi.org/10.1609/aimag.v31i3.2303.

7 Dvid Ferrucci and Adam Lally. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004. doi:10.1017/S1351324904003523.

8 Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346, 1990. URL: http://www.sciencedirect.com/science/article/pii/0167278990900876, doi:https://doi.org/10.1016/0167-2789(90)90087-6.

9 Harding Jenny. Geo-ontology concepts and issues. Technical report, Ilkley UK, September 2002.

10 Jerrold J. Katz and Jerry A. Fodor. The structure of a semantic theory. *Language*, 39(2):170–210, 1963. URL: http://www.jstor.org/stable/411200.

11 Rojanee Khummongkol and Masao Yokota. Computer simulation of human–robot interaction through natural language. *Artificial Life and Robotics*, 21(4):510–519, Dec 2016. URL: https://doi.org/10.1007/s10015-016-0306-5, doi:10.1007/s10015-016-0306-5.

12 Parisa Kordjamshidi, Marie-Francine Moens, and Martijn van Otterlo. Spatial role labeling: Task definition and annotation scheme. pages 413–420. Calzolari, Nicoletta, European Language Resources Association (ELRA), 2010.

13 Ronald W Langacker. *Concept, image, and symbol: the cognitive basis of grammar*. Berlin – New York : Mouton de Gruyter, 1991.

14 E Leisi. Der waltinhalt: Seine struktur im deutschen und englischen. Technical report, Quelle and Meyer, Heidelberg, 1996.

15 H J Levesque, editor. *The Winograd Schema Challenge*, Palo Alto CA, March 2011. AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.

16 Gordon D. Logan and Daniel D. Sadler. chapter A computational analysis of the apprehension of spatial relations., pages 493–529. Language, speech, and communication. The MIT Press, Cambridge, MA, US, 1996.

17 George A. Miller and Philip N. Johnson-Laird. *Language and perception.* Belknap Press, Cambridge, MA, England, 1976.

18 Peter Roget. *Thesaurus of English Words and Phrases.* Longman Dictionaries, 1975.

19 Roger C. Schank and Robert P. Abelson. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Lawrence Erlbaum, Oxford, England, 1977.

20 Roger C. Schank and Larry Tesler. A conceptual dependency parser for natural language. In *Proceedings of the 1969 Conference on Computational Linguistics*, COLING '69, pages 1–3, Stroudsburg, PA, USA, 1969. Association for Computational Linguistics. URL: https://doi.org/10.3115/990403.990405, doi:10.3115/990403.990405.

21 John Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Jan 2000.

22 Leonard Talmy. *Toward a Cognitive Semantics.* MIT Press, 2000.

23 Nigel J.T. Thomas. Mental imagery. *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*, 2018. URL: https://plato.stanford.edu/archives/spr2018/entries/mental-imagery/.

24 Clive Thompson. What is i.b.m.'s watson? New York Times Magazine, June 2010.

25 David E. Wilkins and Karen L. Myers. A Common Knowledge Representation for Plan Generation and Reactive Execution. *Journal of Logic and Computation*, 5(6):731–761, 12 1995. URL: https://doi.org/10.1093/logcom/5.6.731, arXiv:http://oup.prod.sis.lan/logcom/article-pdf/5/6/731/6322894/5-6-731.pdf, doi:10.1093/logcom/5.6.731.

26 Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1 – 191, 1972. URL: http://www.sciencedirect.com/science/article/pii/0010028572900023, doi:https://doi.org/10.1016/0010-0285(72)90002-3.

27 Terry Winograd. Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial Intelligence*, 170(18):1256 – 1258, 2006. Special Review Issue. URL: http://www.sciencedirect.com/science/article/pii/S0004370206000920, doi:https://doi.org/10.1016/j.artint.2006.10.011.

28 Masao Yokota. An approach to natural language understanding based on a mental image model. In *NLUCS*, pages 22 – 31, 2005.

29 Masao Yokota. Towards a universal knowledge representation language for ubiquitous intelligence based on mental image directed semantic theory. In Jianhua Ma, Hai Jin, Laurence T. Yang, and Jeffrey J.-P. Tsai, editors, *Ubiquitous Intelligence and Computing*, pages 1124–1133, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

30 Masao Yokota. Towards a universal language for distributed intelligent robot networking. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3304–3309, Oct 2006. doi:10.1109/ICSMC.2006.384628.

31 Masao Yokota. Aware computing in spatial language understanding guided by cognitively inspired knowledge representation. *Appl. Comp. Intell. Soft Comput.*, 2012:5:5–5:5, January 2012. URL: http://dx.doi.org/10.1155/2012/184103, doi:10.1155/2012/184103.

32 Masao Yokota. *Natural language understanding and cognitive robotics.* CRC press, in press.

Table 3 Questions about I1 - I3 and answers by CMS

Questions about I1 (Tom was with the book in the bus running from Town to University.)	Answers by CMS
Q1: What ran?	A1: bus
Q2: What was in the bus?	A2: Tom, book
Q3: What traveled from Town to University?	A3: Tom, bus, book
Q4: Did the bus carry Tom from Town to University?	A4: yes
Q5: Did the bus move Tom from Town to University?	A5: yes
Q6: Did the bus carry the book from Town to University?	A6: yes
Q7: Did the bus move the book from Town to University?	A7: yes
Q8: Did Tom carry the book from Town to University?	A8: yes
Q9: Did Tom move the book from Town to University?	A9: yes
Questions about I2 (Tom was with the book in the car driven from Town to University by Mary.)	Answers by CMS
Q1: What was driven?	A1: car
Q2: What was in the car?	A2: Mary, Tom, book
Q3: What traveled from Town to University?	A3: Mary, Tom, book, car
Q4: Did the car carry Tom from Town to University?	A4: yes
Q5: Did the car move Tom from Town to University?	A5: yes
Q6: Did the car carry the book from Town to University?	A6: yes
Q7: Did the car move the book from Town to University?	A7: yes
Q8: Did Tom carry the book from Town to University?	A8: yes
Q9: Did Tom move the book from Town to University?	A9: yes
Q10: Did Mary carry the car from Town to University?	A10: yes
Q11: Did Mary carry the book from Town to University?	A11: yes
Q12: Did Mary carry Tom from Town to University?	A12: yes
Questions about I3 (Tom kept the book in a box before he drove the car from Town to University with the box.)	Answers by CMS
Q1: What traveled from Town to University?	A1: Tom, book, box, car
Q2: Did the car carry Tom from Town to University?	A2: yes
Q3: Did the car carry the box from Town to University?	A3: yes
Q4: Did the car carry the book from Town to University?	A4: yes
Q5: Did Tom carry the car from Town to University?	A5: yes
Q6: Did Tom carry the box from Town to University?	A6: yes
Q7: Did Tom carry the book from Town to University?	A7: yes
Q8: Did the box carry the book from Town to University?	A8: yes