Multimedia Description Language and Its Application to Intelligent Cross-media Referencing Systems

Daisuke Hironaka, Seio Oda, Valbone Barolli, Genci Capi, Masao Yokota

Faculty of Information Engineering
Fukuoka Institute of Technology,
E-mail: yokota@fit.ac.jp

Abstract

In general, it is very difficult for people to exploit necessary information from the immense multimedia contents over the WWW. It is still more difficult to search for desirable contents by queries in different media, for example, text queries for pictorial contents. In this case, intelligent systems facilitating cross-media reference are very helpful. Mental Image Directed Semantic Theory (MIDST) has proposed a cognitively motivated methodology for intermediate knowledge representation. This paper presents a formal language for describing multimedia contents, L_{md} , whose syntax and semantics are based on MIDST and its application to cross-media reference between linguistic and pictorial expressions of space and time, which we think will serve as a good basis for more intelligent networking.

1. Introduction

The need for more human-friendly intelligent systems has been brought by rapid increase of aged societies, floods of multimedia information over the WWW, development of robots for practical use, and so on.

For example, it is very difficult for people to exploit necessary information from the immense multimedia contents over the WWW. It is still more difficult to search for desirable contents by queries in different media, for example, text queries for pictorial contents. In this case, intelligent systems facilitating cross-media reference are very helpful.

In order to realize these kinds of intelligent systems, we think it is needed to develop such a computable knowledge representation language for multimedia contents that should have at least a capability of representing spatio-temporal events that people perceive in the real world. In this research area, it is most conventional that conceptual contents conveyed by information media such as languages and pictures are represented in computable forms independent of each other and translated via 'transfer' processes so called which are often very specific to task domains [8], [9], [10].

Yokota, M. et al have proposed a semantic theory for natural languages so called 'Mental Image Directed Semantic Theory (MIDST)' [2]. In MIDST, word concepts are associated with omnisensual mental images of the external or physical world and are formalized in an intermediate language L_{md}, based on first-order predicate logic while the other knowledge description schema [3], [4] are too linguistic (or English-like) to formalize omnisensual mental images.

MIDST has been implemented on several types of computerized intelligent systems [1], [5], and there is a feedback loop between them for their mutual refinement, unlike other similar theories [6], [7].

This paper presents the formal language L_{md} and its application to cross-media translation and question-answering (Q-A) between linguistic and pictorial expressions of space and time.

2. Multimedia description language, L_{md}

The language L_{md} is employed for many-sorted first-order predicate logic containing one special predicate with five types of terms. The most remarkable feature of L_{md} is its capability of formalizing both temporal and spatial event concepts on the level of human sensations while the other similar knowledge representation languages are designed to describe the logical relations among conceptual primitives such as words [3], [4].

2.1 Atomic locus formula

MIDST treats word meanings in association with mental images, not limited to visual but omnisensual, modeled as "Loci in Attribute Spaces" [2]. An attribute space corresponds with a certain measuring instrument just like a barometer, a map measurer or so and the loci represent the movements of its indicator.

A general locus is to be articulated by "Atomic Locus" formalized as the expression (1). This is a formula in many-sorted first-order predicate logic, where "L" is a predicate constant with five types of terms: "Matter" (at 'x' and 'y'), "Attribute Value" (at 'p' and 'q'),. "Attribute" (at 'a'), "Event Type" (at 'g') and "Standard" (at 'k').

$$L(x,y,p,q,a,g,k) \qquad (1)$$

The formula is called "Atomic Locus Formula" whose first and second arguments are sometimes referred to as 'Event Causer ' and 'Attribute Carrier ', respectively.

The interpretation of the expression (1) is intuitively given as follows, where "matter" refers to "object " or "event".

"Matter 'x' causes Attribute 'a' of Matter 'y' to keep (p=q) or change (p ≠ q) its values temporally (g=Gt) or spatially (g =Gs) over a time-interval, where the values 'p' and 'q' are relative to the standard 'k'."

When g=Gt and g=Gs, the locus indicates monotonous change or constancy of the attribute in time domain and that in space domain, respectively. The former is called a temporal event and the latter, a spatial event.

For example, the motion of the 'bus' represented by S1 is a temporal event and the ranging or extension of the 'road' by S2 is a spatial event whose meanings or concepts are formalized as (2) and (3), respectively, where the attribute is "physical location " denoted by 'A12'.

(S1) The bus runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A12,Gt,k)\wedge bus(y) \qquad (2)$$

(S2) The road runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A12,Gs,k)\wedge road(y) \qquad (3)$$

2.2 Tempo-logical connectives

The expression (4) is the conceptual description of the English word "fetch", implying such a temporal event that 'x1' goes for 'x2' and then comes back with it, where 'Π'and '•' are instances of the tempo-logical connectives, 'SAND' and 'CAND', standing for "Simultaneous AND" and "Consecutive AND", respectively. In general, a series of atomic locus formulas with such connectives is called 'Locus formula'.

$$\begin{aligned}&(\exists x1,x2,p1,p2,k)\ L(x1,x1,p1,p2,A12,Gt,k)\\&\bullet (L(x1,x1,p2,p1,A12,Gt,k)\\&\Pi\ L(x1,x2,p2,p1,A12,Gt,k))\wedge x1\neq x2\wedge p1\neq p2 \qquad (4)\end{aligned}$$

Furthermore, a very important concept called 'Empty Event (EE)' and symbolized as 'ε' must be introduced. An EE stands for nothing but for time collapsing and is explicitly defined as (5) with the attribute 'Time Point (A34)'. It is essentially significant for the MIDST that ***every temporal relation can be represented by a combination of Empty Events, SANDs and CANDs***. For example, (6) represents 'X_1 during X_2'.

$$\varepsilon \Leftrightarrow (\exists x,t,p,q,g,k)\ L(x,t,p,q,A34,g,k)\wedge time(t) \qquad (5)$$

$$(\varepsilon_1 \bullet X_1 \bullet \varepsilon_2)\ \Pi\ X_2 \qquad (6)$$

2.3 Attributes and standards

The attribute spaces for humans correspond to the sensory receptive fields in their brains. At present, about 50 attributes concerning the physical world have been extracted exclusively from English and Japanese words as shown in Table 1. They are associated with all of the 5 senses (i.e. sight, hearing, smell, taste and feeling) in our everyday life while those for information media other than languages correspond to limited senses. For example, those for pictorial media, as a matter of course, associate limitedly with the sense 'sight' as are marked with '*' in Table 1.

Correspondingly, six categories of standards shown in Table 2 have been extracted that are assumed necessary for representing values of each attribute in Table 1. In general, the attribute values represented by words are relative to certain standards as explained briefly in Table 2.

2.4 Event concept description

Event concepts are described as locus formulas according to the hypothesis as follows.

(Hypothesis 1) *Every event, a temporal, a spatial or a mixed one is perceived through a certain movement of the focus of attention of the observer (FAO) over the matters appearing in it and such perception is associated with temporal loci in attribute spaces reflecting the FAO movement. Therefore, the concept of an event is a generalization of such perceptions.*

Considering such sentences as S3 and S4, the underlined parts imply some events neglected in time and in space, so called, 'Temporal Empty Event (TEE)' and 'Spatial Empty Event (SEE)', respectively.

The concepts of S3 and S4 are represented as (7) and (8), where TEE and SEE are symbolized as 'ε_t' and 'ε_s', respectively. Matter terms placed at Attribute Values or Standard represent their values at the time for simplicity. The attribute 'A15' means 'Trajectory' and (8) can refer such a spatial event as depicted in Fig.1 while (7) should be interpreted into animation. The special symbol '_' defined by (9) is often used instead of the variable bound by an existential quantifier especially when it has little significance. For example, 'Event Causer' in the event referred to by an intransitive verb such as 'run' in S1 or S2.

Table 1. A part of attributes extracted from linguistic expressions.
The properties "S" and "V" represent "scalar" and "vector", respectively.

Code	Attribute [Property]	Linguistic expressions for attribute values.
*A01	PLACE OF EXISTE NCE [V]	He is in Tokyo. The accident happened in Osaka.
*A02	LENGTH [S]	The stick is 2 meters long (in length).
	....	
*A09	AREA [S]	The crop field is 10 square miles.
*A10	VOLUME [S]	The box 10 cubic meters.
*A11	SHAPE [V]	The cake is round.
*A12	PHYSICAL LOCATION [V]	Tom moved to Tokyo.
*A13	DIRECTION [V]	The box is to the left of the chair.
*A14	ORIENTATION [V]	The door faces to south.
*A15	TRAJECTORY [V]	The plane circled in the sky.
*A16	VELOCITY [S]	The boy runs very fast.
*A17	DISTANCE [S]	The car ran ten miles.
A18	STRENGTH OF EFFECT [S]	He is very strong.
	....	
*A32	COLOR [V]	The apple is red. Tom painted the desk white.
A33	INTERNAL SENSATION [V]	I am very tired.
A34	TIME POINT [S]	It is ten o'clock.
	....	

(S3) The *bus* runs 10km straight east from A to B, and
after a while,
at C it meets the street with the sidewalk.

(S4) The *road* runs 10km straight east from A to B, and
after a while,
at C it meets the street with the sidewalk.

$$\begin{aligned}&(\exists x,y,z,p,q)\\ &\quad (L(\_,x,A,B,A12,Gt,\_)\ \Pi\ L(\_,x,0,10km,A17,Gt,\_)\\ &\quad \Pi\ L(\_,x,Point,Line,A15,Gt,\_)\\ &\quad \Pi\ L(\_,x,East,East,A13,Gt,\_))\\ &\quad \bullet\ \varepsilon_t \bullet (L(\_,x,p,C,A12,Gt,\_)\ \Pi\ L(\_,y,q,C,A12,Gs,\_)\\ &\quad \Pi\ L(\_,z,y,y,A12,Gs,\_))\\ &\quad \wedge bus(x) \wedge street(y) \wedge sidewalk(z) \wedge p \neq q\end{aligned} \tag{7}$$

$$\begin{aligned}&(\exists x,y,z,p,q)\\ &\quad (L(\_,x,A,B,A12,Gs,\_)\ \Pi\ L(\_,x,0,10km,A17,Gs,\_)\\ &\quad \Pi\ L(\_,x,Point,Line,A15,Gs,\_)\\ &\quad \Pi\ L(\_,x,East,East,A13,Gs,\_))\\ &\quad \bullet\ \varepsilon_s \bullet (L(\_,x,p,C,A12,Gs,\_)\ \Pi\ L(\_,y,q,C,A12,Gs,\_)\\ &\quad \Pi\ L(\_,z,y,y,A12,Gs,\_))\\ &\quad \wedge road(x) \wedge street(y) \wedge sidewalk(z) \wedge p \neq q\end{aligned} \tag{8}$$

$$L(\ldots,\_,\ldots) \leftrightarrow \exists x L(\ldots,x,\ldots) \tag{9}$$

The expression (7) shows a mixed event where temporal and spatial ones coexist. Not limited to this expression, it is generally postulated that any event coexists with a temporal event of the attribute 'place (A01)' in the world, which can be formalized as expression (10), where X denotes a locus formula, implying that any matter appearing in an event occupies some place in the world during the event.

$$(\forall x)(X(x). \supset .(\exists y,p,k)X(x)\,\Pi L(y,x,p,p,A01,Gt,k)) \tag{10}$$

It is psychologically pointed out that more than one matters can be sensed simultaneously in the 'field of attention' while the 'focus of attention' i.e. FAO can be taken on each matter in a time-sharing way. The usages of CANDs and SANDs here reflect this psychological fact.

For example, the event referred to by S5 can be represented as the expression (11).

(S5) The UFO changed its shape from triangle to square and then to pentagon while the ray changed its color from red to orange and then to yellow.

$$\begin{aligned}&(L(\_,x,Triangle,Square,A11,Gt,k)\\ &\bullet L(\_,x,Square,Pentagon,A11,Gt,k))\\ &\Pi\ (L(\_,y,Red,Orange,A32,Gt,k)\\ &\bullet L(\_,y,Orange,Yellow,A32,Gt,k)) \wedge ray(x) \wedge UFO(y)\end{aligned} \tag{11}$$

As easily understood by this example, every continuous locus is articulated at the outstanding points where its attribute value changes abruptly or so and is formalized as a sequence of atomic formulas connected by CANDs.

3. Cross-media reference

3.1 Functional requirements

The authors have considered that systematic cross-media reference must be realized on the basis of systematic cross-media translation and which in turn must have such functions as follows.

(F1) To translate source representations into target ones as for contents describable by both source and target media. For example, positional relations between/among physical objects such as 'in', 'around' etc. are describable by both linguistic and pictorial media.

Table 2. Standards of attribute values.

Categories of standards	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*'.

(F2) To filter out such contents that are describable by source medium but not by target one. For example, linguistic representations of 'taste' and 'smell' such as 'sweet candy' and 'pungent gas' are not describable by usual pictorial media although they would be seemingly describable by cartoons, etc.

(F3) To supplement default contents, that is, such contents that need to be described in target representations but not explicitly described in source representations. For example, the shape of a physical object is necessarily described in pictorial representations but not in linguistic ones.

(F4) To replace default contents by definite ones given in the following contexts. For example, in such a context as "There is a box to the left of the pot. The box is red. ...", the color of the box in a pictorial representation must be changed from default one to red.

3.2 Formalization

According to MIDST, any content conveyed by an information medium is assumed to be associated with the loci in certain attribute spaces and in turn the world describable by each medium can be characterized by the maximal set of such attributes. This relation is conceptually formalized by the expression (12), where Wm, Am_i, and F mean 'the world describable by the information medium m', 'an attribute of the world', and 'a certain function for determining the maximal set of attributes of Wm', respectively.

$$F(Wm)=\{Am_1, Am_2, \ldots, Am_n\} \quad (12)$$

Considering this relation, cross-media translation is one kind of mapping from the world describable by the source medium (ms) to that by the target medium (mt) and can be defined by the expression (13).

$$Y(Smt)=\psi(X(Sms)), \quad (13)$$

where

Sms: the maximal set of attributes of the world describable by the source medium ms ,

Smt: the maximal set of attributes of the world describable by the target medium mt,

$X(Sms)$:a locus formula about the attributes belonging to Sms,

$Y(Smt)$: a locus formula about the attributes belonging to Smt ,

and

ψ : the function for transforming X into Y, so called, 'Locus formula paraphrasing function'.

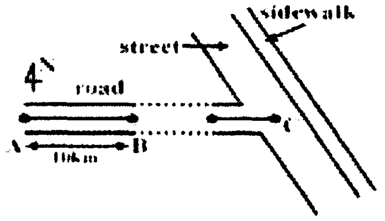


Fig 1. Illustration of the expression (8).

Table 3. Fundamental specifications for text-to-picture translation.

	Categories of media	Maximal sets of attributes - Sms and Smt	Categories of standards	UI-Devices
Source medium	Natural language texts	Sms = All the attributes of Table 1	All of Table 2	Keyboard
Target medium	2-D drawings	Smt = Ones marked by * in Table 1	Rigid Standard	CTV monitor

Table 4. Attribute paraphrasing rules for text-to-picture translation.

APRs	Correspondences of attributes (Text : Picture)	Value conversion schema (Text → Picture)	Interpretations of the schema
APR-01	A12 : A12	$p \to p'$	'position' into 2D coordinates (within the display area).
APR-02	{A12, A13, A17} : A12	$\{p, d, l\} \to p'+l'd'$	{'position', 'direction', 'distance'} into 2D coordinates.
APR-03	{A11, A10} : A11	$\{s, v\} \to v's'$	{'shape', 'volume'} into a set of outlines of the object.
APR-04	A32 : A32	$c \to c'$	'color' into 2D coordinates of the color solid.
APR-05	{A12, A44} : A12	$\{p_a, m\} \to \{p_a', p_b'\}$	{'position', 'topology'} into a pair of 2D coordinates.

The function ψ is designed to clear all the functions F1-F4 by inference processing at the level of locus formula representation.

3.2 Locus formula paraphrasing function ψ

In order to satisfy the function F1, a certain set of '***Attribute paraphrasing rules (APRs)***', so called, are defined ***at every pair of source and target media*** (See Section 4). The function F2 is satisfied by detecting locus formulas about ***the attributes without any corresponding APRs*** from the content of each input representation and replacing them by ***empty events***.

For the function C3, ***default reasoning*** is employed. That is, such an inference rule as defined by the expression (14) is introduced, which states if ***X is deducible and it is consistent to assume Y then conclude Z***.

This rule is applied typically to such instantiations of X, Y and Z as specified by the expression (15) which means that the indefinite attribute value 'p' with the indefinite standard 'k' of the indefinite matter 'y' is substitutable by the constant attribute value 'P' with the constant standard 'K' of the definite matter '$O\#$' of the same kind 'M'.

$$X \circ Y \rightarrow Z \qquad (14)$$

$$\{\ X / (L(x,y,p,p,A,G,k) \wedge M(y)) \wedge (L(z,O\#,P,P,A,G,K) \wedge M(O\#)),$$
$$Y / p=P \wedge k=K,$$
$$Z /\ L(x,y,P,P,A,G,K) \wedge M(y)\ \} \qquad (15)$$

The satisfaction of the function C4 is realized quite easily by ***memorizing the history of applications of default reasoning***.

3.3 Translation between text and picture

Cross-media translation between text and picture is basically specified by the items in Table 3 which concern the categories, maximal sets of attributes, standards for attribute values of the source and target media and the user-interface devices. It is most remarkable that all the standards employed for target pictorial representation are specific to the performances of the color television monitor (CTV monitor), e.g. the size of the screen, belonging to the category 'Rigid Standard'.

Five kinds of APRs for this case are shown in Table 4 where $p,s,c,...$and $p',s',c',...$are linguistic expressions and their corresponding pictorial expressions of attribute values, respectively. Further details are as follows:

(1) APR-02 is used especially for a sentence such as "The box is 3 meters to the left of the chair." The symbols p, d and l correspond to 'the position of the chair', 'left' and '3 meters', respectively, yielding the pictorial expression of 'the position of the box', namely, "$p'+l'd'$ ".

(2) APR-03 is used especially for a sentence such as "The pot is big." The symbols s and v correspond to 'the shape of the pot (default value)' and 'the volume of the pot ('big')', respectively. In pictorial expression, the shape and the volume of an object is inseparable and therefore they are represented only by the value of the attribute 'shape', namely, $v's'$.

(3) APR-05 is used especially for a sentence such as "The cat is under the desk." The symbols p_a, p_b and m correspond to 'the position of the desk', 'the position of the cat' and 'under' respectively, yielding a pair of pictorial expressions of the positions of the two objects.

4. Implementation on IMAGES-M

The intelligent system IMAGES-M, still under construction, is intended to facilitate integrated multimedia information understanding, including cross-media reference. IMAGES-M consists of such seven major modules as follows:

(1) Text Processing Unit (TPU),
(2) Speech Processing Unit (SPU),
(3) Picture Processing Unit (PPU),
(4) Animation Processing Unit (APU),
(5) Sensory Data Processing Unit (SDPU),
(6) Inference Engine (IE), and
(7) Knowledge Base (KB).

The collaboration of TPU and PPU/APU for text-to-still-picture / text-to-animation translation is carried out as follows:

(STEP-1) TPU takes in a text and translates it into a conceptual description, so called, 'Text Meaning Representation (TMR)', where syntax and meaning dictionaries in KB are utilized through IE.

(STEP-2) IE, using APRs stored in KB, deduces from the TMR a conceptual description adaptive for picture/animation, so called, 'Picture Meaning Representation (PMR)' which reflects all the functional constraints of the output device, color TV monitor. For a very trivial example, such a device can display 'a red candy', but not 'a *sweet* candy'. That is, PMR does not contain any locus formula of taste, smell, or so but of vision as shown in Table 1.

(STEP-3) PPU/APU interprets the PMR as a still picture / an animation. The choice between PPU and APU depends on the verb concepts contained in the text. For example, the PMRs for S3 and S4 are sent to APU and PPU, respectively.

The reverse process, namely, picture-to-text translation has been also realized. Figures 2 and 3 are examples of text-to-picture and picture-to-text translation, respectively.

Cross-media Q-A is also executable by inference on TMRs and PMRs. Figure 4 shows an example of cross-media Q-A, where the headers 'H' and 'S' represent 'Human user' and 'System (=IMAGES-M)', respectively.

5. Discussion and conclusion

The cross-references between texts in several languages (Japanese, Chinese, Albanian and English)·and pictorial patterns such as maps were almost successfully implemented on our intelligent system IMAGES-M with the functions F1-F4. This leads to the conclusion that employment of atomic locus formulas has made the logical expressions of event concepts remarkably computable and has proved to be very adequate to systematize cross-media reference because of their medium-freeness. We think that this research work can serve as a basis for building more intelligent networking systems.

The problems for the future of our project are as follows:

(P1) Augmentation of text understanding for dissolving ambiguity and vagueness.

(P2) Augmentation of picture generation for 3D graphics.

(P3) Continuation of conceptual analysis and description of words.

Karrigia eshte 3m ne te djathte te vazos..

猫は椅子の 1 m 下にいる.

Macja eshte e kuqe..
the small box is 1m to the left of the chair..
the big blue lamp is 2m above the pot..

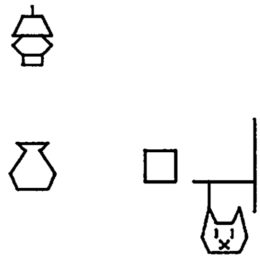
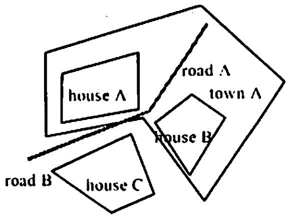


Fig 2. Text-to-picture translation



The house A is in the town A.
The house B is in the town A.
The house A is 174 m
to the upper left of the house B.
The road B is between
the house A and the house C.
⋮

Fig 3. Picture-to-text translation.

H: ?猫 是 紅的 (Is the cat red ?)
S: 是 (yes)
H: ?何が椅子と花瓶の間にある
(What is between the chair and the flower-pot ?)
S: 箱 (box)
H: Is the box between the cat and the pot ?
S: NO
H: Eshte kutia midis maces dhe llampes ?
(Is the box between the cat and the lamp ?)
S: PO (yes)

Fig 4. Q-A for the picture in Fig.2.

References

[1] D. Hironaka, S. Oda, K. Ryu & M. Yokota : "Mutual Conversion of Sensory Data and Texts by an Intelligent System IMAGES-M", Proc. of the 8th International Symposium on Artificial Life and Robotics (AROB '03), pp.141-144, 2003.

[2] M. Yokota, et al: "Mental-image directed semantic theory and its application to natural language understanding systems", Proc. of NLPRS'91, pp.280-287, 1991.

[3] J.F. Sowa: Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[4] G.P. Zarri: "NKRL, a Knowledge Representation Tool for Encoding the 'Meaning' of Complex Narrative Texts", Natural Language Engineering - Special Issue on Knowledge Representation for Natural Language Processing in Implemented Systems, 3,pp.231-253, 1997.

[5] S.Oda, M.Oda & M.Yokota : "Conceptual Analysis Description of Words for Color and Lightness for Grounding them on Sensory Data", Trans.of JSAI,16-5-E,pp.436-444, 2001.

[6] R.W. Langacker : Concept, Image and Symbol, Mouton de Gruyter, Berlin/New York, 1991.

[7] G.A. Miller & P.N.: Johnson-Laird : Language and Perception, Harvard University Press, 1976.

[8] A.Yamada, et.al.: Reconstucting spatial image from natural language texts, in Proc. of Coling 90, Nantes, 1992

[9] P.Olivier & J.Tsujii: A Computational View of the Cognitive Semantics of Spatial Expressions, Proc. of ACL 94, Las Cruces, 1994.

[10] G. Adorni, M. Di Manzo, & F. Giunchiglia. Natural Language Driven Image Generation. Proc. of COLING 84, pp. 495-500, 1984