A PLOT UNDERSTANDING SYSTEM ON REFERENCE TO BOTH IMAGE AND LANGUAGE

Norihlro Abe, Itsuya Soga, Saburo Tauji

Department of Control Engineering
Faculty of Engineering Science
Osaka University
Toyonaka Osaka Japan

ABSTRACT

In this paper, a system is described that can understand a plot of a story on reference to both image and linguistic information. As input, a series of line drawings with colors and narrations in English concerning to these drawings are given to the system.

It searches the objects suggested to be in the scene by the narrations, finding relations among them, making the world model by using its world knowledge. Its refernce to those drawings makes it easy for system to analize complicated structures in the narration sentences such as those of prepositions, and guides the process reasoning about the CD representation using rules and demons. At the end of this paper, a result on QA will be shown.

1. Introduction

Recently the research on Artificial Intelligence has become widely practiced, and there have been many papers on story understanding. They are roughly categorized into two classes. The first one corresponds to researches done by Schank] Lehnert,[2] Wilensky[3] and Carbonell[4] which infer causal relation among things using so-called common-sense such as scripts, goals or human traits etc. They are quite interesting in respect of the points that they contribute to better comprehension of the flow of events in the story. AS

The other is on picture processings; its goal is to extract some meaningful actions by observing a sequence of movements in images. This research is also interesting because it may suggest any insight as to how we can recognize higher concept on actions from the view. But its consequence is not necessarily satisfactory as it is difficult to capture action with the help of movements in the scene alone. The reason for this can be well understood considering our cognitive methodology.

In our life, we utilize many sorts of knowledge sources to capture things surrounding us. We cannot well understand TV-movies without any narration, and an infant child cannot follow the story without pictures. Thus, computer program should positively refer to these sources.

Concerning to this point, Waltz[7] has written a paper pointing out that pictorial knowledge often makes it easier to infer the relation not easily deducible from linguistic knowldge. But using the model as a default one, it has been not necessarily useful for the recognition of a real situation being at hand, because the verification has not been attempted if the predicted things can be well matched to those in the scene.

Seeing these progresses on AI, it is convinced that the time has come when an integrated system that can understand both image and language is implemented. In this paper, by referring to both language and image data or knowledge, it is shown that a program can appreciate events which occur in the story. The method taken in this paper to correlate linguistic data with pictorial one can be considered relevant not only to the cognitive models required to simulate our epistemology, but also to the practical objective that picture processing programs should be used in more flexible styles by using natural language interface to access a proper portion of images.

2. The overview of this system

In this system, at first a series of simple line drawings with color informations and narrations corresponding to these images are given to the system as its input. Then the system tries to do reasoning about the plot of the input story referring to such image and language data.

2.1 A cross reference to visual and linguistic Information

Though this system is partioned into an image data processing part and a language interpretation part, they do not work with their objects for themselves but interact with each other to take full advantage of the consequence brought about from the other part. This mutual reference leads the system to a better cognitive level than in cases they are used alone. It appears that such a behavior of the system well resembles to our daily behavior.

When we see a certain scene, it is sure that the entire information involved in that scene ends up on our sense organ, for all that there are cases where some of their features or even their entities are not explicitly realized unless they are specifically interesting for us. In other words, our cognitive mechanism is apt to cause something to be evaluated in order to keep them in view as a highly processed data if they are interesting, but something to be ignored because they deserve no attention. But this does not mean that such insignificant things entirely slip out of our memory but that they are held implicitly in forms of an unconscious low level data, more concretely in a coordinate value. As one of such triggers that focuses an attention on objects, an indication by language can be regarded.

For this reason, in the course of the inspection of objects in given scenes, a computer program should positively use language data closely related to those images, making it possible to restrict its search domain to a small plausible region, and leaving low level data intact in the form of coordinate values until they are needed.

••Visual information is regarded as an essential one for us to capture the complex relations among things, with respect to thier locations. For example, when we ask someone to show the way to station, if he illustrates the way showing a map, it would be easier for us to understand what he says. The computer program should utilize such a pictoral aid concurrently with referring to linguistic knowledges, and then it is easy to analyze the structure of sentences such as those between prepositional groups.

•* It is needless to say that the precise comprehension of the meaning of sentences or words with various kinds of sense requires some contextual informations and so-called common-sense knowledge, and there are many sorts of such contextual information. Visual information can also be thought to be a member of such context builders. Considering this :

He takes the apple in the box.

After reading this fragment alone, it is impossible to decide whether "take" means the action "eat" or "get" Even in such a case, the appreciation of informations illustrating the scene leads the sentence analyzer to eas> settlement of such a problem.

2.2 The assumption of the system

So far, we have written as if all things worked well by taking full advantages of both informations, in reality, however, several assumptions must be set up as in the followings.

(1) A situation that a line drawing illustrates corresponds to the time when the affairs described in the narration is completed in the aspectual sense.

(2) As a rule, unless objects are explicitly mentioned in the narration even if they are identifiable in the given scene, the program does not look for them in the scene and therefore they are not remembered in the *forms of* assertions as to be in the scene.

Why the maneuver like this is taken here must be described. We have already mentioned the reason in 2.1 from one perspective, that is "interesting". And from another perspective, if such insignificant objects and all the relations among them were memorized in the memory, the combinatorial explosion on the management of such relations would be unavoidable, because a slight change concerning to one of such insignificant things would cause drastic reformation. To put it more concretely, if a person standing in a situation consisting of some rocks moves toward one of them, it is not sufficient to make an alteration only to the relation between the person and the rock. All relations between the person and the other objects in the scene must also be changed, because that person surely approached one of them, but simultaneously went away from the other of them. In case where a situation becomes more complicated, doing such works would make the system unmanageable.

(3) Objects in the scene are looked for in terms of their colors and relations among their sub-parts. Any rotational movements around any axis is assumed not to happen for the simplicity. The reason why these assumptions are required is almost clear considering the current state of art on image processing techniques. As this research is an experimental attempt referring to two sorts of knowledge sources, such topics are not taken into considerations.

In the followings, beginning with a data structure used in this system, moving through concrete explanation on programs, and ending with their implementation results.

3. Knowledge representation end dete structure

It is well known that knowledge is indispensable to our daily life, but this is also true for a computer when we let him appreciate an environment surrounding him. In this regard, the world knowledge, object models for an object Identification, and dictionaries for language analysis are given to this system as its apriori knowledges, which are of vital importance to make inference about a plot of the given story. This reasoning process involves a generation of the description representing their causal chains among events, a construction of the world model in terms of their locations and that of the frame model containing several sorts of properties of objects, and these consequences will be used by our question answering programs mentioned in 5.2. In this section, the data structures will be shown.

3.1 World Knowledge

The implicit knowledge, which we feel we have, explaining various events in the world are built into the system in forms of the semantic network, with a hierarchical structure described in terms of general/specific relations. The network consists of objects or nodes representing their concept and linkages semantically linking them. The node is in the form of frame structure, which have several number of slots for properties or attributes that object will have, and the lowest node in the tree corresponds to the specific object as shown in Fig.l. The linkage between nodes has a label-value pair which illustrates a relation with its score holding between those frames. As is the case with the ordinary frame system, some functions enable our frame to inherit information from the higher frame by following SIS-A relations(see Fig.l).

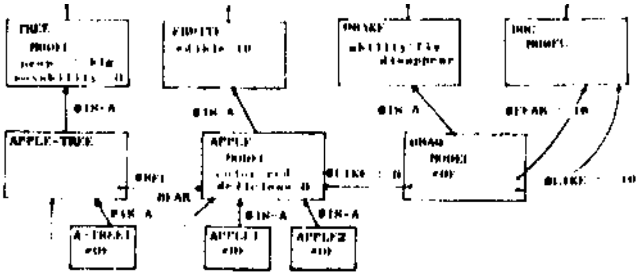


Fig.1 The world knowledge.

3.2 Models for object Inspection

With respect to the inspection of object in a scene, the following two points are assumed.

(1) We permit the line drawing given as input to have some deep information with respect to the actor's movement toward fore and aft directions in the approximate sense.

(2) Models are designed to make it easy to describe objects with respect to the mutual relations among their subparts and their colors, and their precise shape is not taken into consideration from the following reason:

When we observe fairly complicated objects in our daily life, we often capture them as the composite set consisting of their elementary subparts, and are not possessed with an awkward idea that they should be matched against their models in a strict sense.

Concerning to the object identification, taking this flexibility Into consideration results in a robust program, that is, less subjective to several kinds of noise, and the program can, therefore, recognize what he sees.

A format of the object models is shown in Fig.2. As shown in this figure, the object to be inspected is framed with a rectangle just fitting the boundaries of the object. Next, this rectangle is divided into 9 sub-regions, where each sub-parts locates and what relations it has among others are described in terms of these sub-regions. The assumption that no rotational movement occurs helps to simplify this framing process, as all of its edges are restricted to vertical or horizontal ones.

Fig.3 shows a model for desks. At top of it, there is a declaration preceded by the symbol *PICT, which says that the main property is a DESK, and its subordinate statements, preceded by *SUBR, follows which means that a breakup operation of it into its subparts.

This breakup process can be repeated so that a hierarchical representation of models are built up. With respect to each statement, the subpart, its color, location and relations among other parts are put in their own places, and yet to each of such statements, a score is appended which is used in a decision whether the matching of object against this model succeeds or not. For an instance, a description (FOOT1 COUT ((R) D) 10) means the fallowings,

• FOOT1 is located in a position shown in Fig.3.

• A relation COUT holds between FOOT1 and DSK which is its superior constituent of FOOT1, where the relation COUT states that the body of FOOT1 is Contignous to that of DSK and that it occupies a region OUTside of DSK.

» If all of these conditions are satisfied, 10 points will be given as its score.

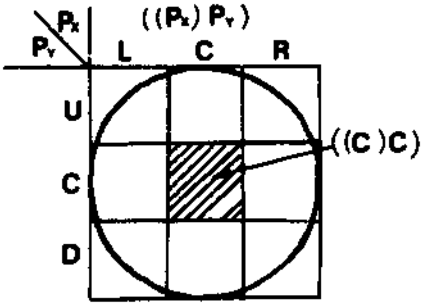


Fig.2 Description of object.

```
(*PICT DESK 25)
       (SUBP DSK 30)

(*SUBR DSK 25)
       (*COL BROWN 10)
       (FOOT1 COUT ((R) D) 10)
       (FOOT2 COUT ((L) D) 10)

(*SUBR FOOT1 10)
       (*COL BROWN 10)

(*SUBR FOOT2 10)
       (*COL BROWN 10)
```

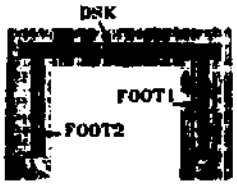


Fig.3 A model of DESK.

3.3 Dictionary

A suitable representaion must be selected so as to make it easy for a program to infere things which might happen and also to make it easy to follow his reasoning when asked some questions. The CD(Conceptual dependency) representation exploited by Schank is adopted by the following reasons:

(1) The CD concept leads an evaluator to straightfoward resolution with respect to its interpretaion of ambiguous sense of words.

(2) The CD makes it easier to reason about the cause-effct relations among events in the story.

In favor of these, the following two dictionaries are used in this system.

3.3.1 Syntactic dictionary

This is a dictionary that is refered by mini-ATN parser which generates a possibility list consisting of all plausible combinations among prepositional groups or fragements of sentence. Lists of ideoms or syntactic roles of words, for some of which the case information, are stored in this dictionary. To date, nearly 200 words have been given their full definitions and more than about 300 words registered into memory in an inperfect form.

3.3.2 CD dictionary

With the help of this dictionary, a surface structure of an input sentence is translated into the *CD* representation. Each verb in this dictionary is given a set of rules evaluated by a pattern-directed invocation. For each of these rules, its premises and conclusions are listed in terms of the CD notion, with a demon to be created and procedures to be executed when this rule does a good job. In other words, each rule is in form of so-called production rule and functions are added to it to realize side-effects that is attended by this action. This demon is required to turn many thing into reality, especially to associate assertions believed to be in the cause-effect relation by going through all assertions, or adding some assertions deducible from this rule when neccessary.

3.4 Object Frame

This is a most specific frame placed in the lowest position in the hierarchical tree, and represents an instance of its higher node. This Object Frame (abbreviated as OF here after) has slots containing its properties, its location and all assertions closely related to this object. These slot structures deserve more explanation because of its idiosyncracy.

As a location of an object will be change when the object moves in a scene, its OF must record the current location in a proper form. For this purpose, the slots must be categorized into two classes, one involving constant properties and the other for dynamic attributes such as its location, state, and actions that have something to do with this object. These dynamic properties are managed under the frame number of the scene, that is, the frame number is used as its leading slot. So the slots in an ordinary sense are secondary and can be used as a key to retrieve information, provided that frame number is known(be careful not to confuse the meaning of these two frames).

4. System Organization

A rough sketch of our system is shown in the followings and we want to put them into perspective.

4.1 Syntactic parser

This is a well known ATN parser, which refers to the syntactic dictionary when necessary. It does so-called the syntactic analysis, and on its completion several parsing trees will be generated in the form of possibility list, each of which shows plausible combinations of prepositional groups or others. For the simplicity, we dismiss this and proceed to next.

4.2 Generation of intermediate notation

After getting parsing trees, they are tested which of them can give the most plausible interpretation for the given sentence by referring to the corresponding image data. To put it more concretiy, the verification is conducted to see whether there exist what are expected to be in the scene and also the relations suggested by this possibility iist hold among them. For example, let the following fragment given along with Fig.4

A cat sees the clock above a box above a desk on the chair.

It is clear that image information contributes to exclude interpretations failig to make sense. In this case, at first object inspector searches objects such as chair and desk, because they are probably sitting on the floor. Then it is easy to see that a noun group "the clock" is modified by two prepositional groups of PREPG1, PREPG2 and that PREPG3 functions as an adverbial phrase. A group of lists illustrated below Fig. 4 shows an intermediate representation for this sentence and inner notations concerning to a location, which are recorded into the memory in terms of symbolic expressions because their relations are mentioned explicitly in English.

Note here that if a sentence is so simple that no confusion arises in its analysis, the object inspector can utilize more directly the results from the sentence analyzer. And it should be also noted that the model inspector can use restriction that the case information of words will give to him. Then the remaining objects are found. In this way, once real relations among them are found in the image, they help to decide relation among phrases.

4.3 Object Inspection

As has already been mentioned, in object models their score are recorded along with their structural constraint, and the effective use of them can enable the pattern matcher to serve as a partial matcher. In this section, how such models are used with the matcher is first shown, and next we provide a simple example to make it easy for readers to be acquainted with matching process.

4.3.1 Utilization of model

Let assume that we are at a verification process if some object can be matched against the model given as in the followings:

```
( (•SUBR A thsc)              <s0>
(•COL B scA2)                 <s1>
(C rel-1 pos-2 scA2)          <s2>
(D rel-3 pos-3 scA2) )        <s3>
```

• Evaluation of this model

Matching is done with respect to some standard reference to the following inequality.

$$\sum (scA \cdot i - xA \cdot i) \geq thsc$$

In this expression, scA-i is to be a score a-priori given to this model and xA-i is to be a real score calculated through the method given as in the followings.

When this inequality holds, after taking all possible components described below into account, this portion of the object is said to be an instance of this fragment of model. The evaluation of the model itself is also done in a similar way to this method, but in that case the topmost statement of the description will be (•PICT A thsc).

* Computation of xA-i

(1) For color statement in<SI>, a color of the object part is checked against the one designated as B in this illustrative model. If they are quite the same one, then set xA-1 to 0, otherwise set it to scA-1. This means that the color information plays a key role in the matching process.

(2) For relational statement in<S2,S3>, the following three components must be taken into account.

[1] Verification of C. Note that a description with respect to C will have also the same style as with the case for A. If the description for C is found successfully by going through the object beings currently inspected, then set y1 to a value given by the following equation, otherwise set it to scA2.

$$y_1 - \frac{scA_2 \sum xC_i}{\sum scC_i}$$

[2] A contiguous relation between A and C must be checked out on this model description. When no discrepancy is found between them, set y2 to 0. If one of their relations is either CIN or COUT and the another is also either of these two specific relations, let y2 be l/4»scA2. Otherwise, that is, if they are quite unlike each other, set y2 to l/2*scA2.

[3] As is the case with [2], a locative relation must be also looked at. No difference between them implies that y3 should be set to 0. If a combination of their positional symbol with respect to A and C is a member

A cat sees the clock above a box
PREPG1
above a desk on the chair.
PREPG2 PREPG3

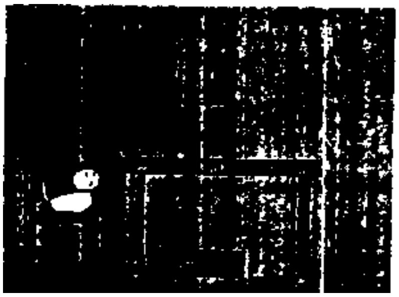
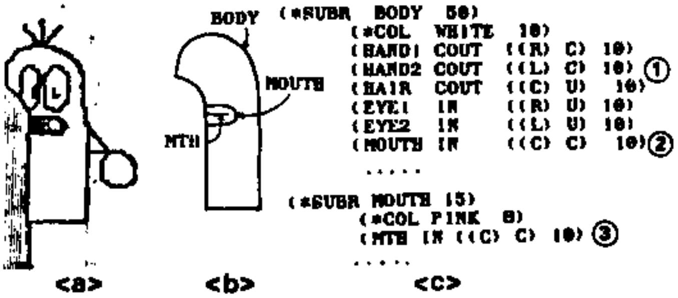


Fig.4 Analysis of prepositional groups.

(CAT1 SEE CLOCK1)
((•POS) CAT1 ON CHAIR1)
((•POS) CLOCK1 ABOVE BOX1)
((•POS) CLOCK1 ABOVE DESK1)

of the set <(U,C) (D,C) (L,C) (C,R)>, then equate y3 to l/4*scA2, otherwise set it to l/2*scA2 At last, xA2 is given a vaiue y1+ y2+ y3.

Looking at its matching process in this light, it is easy for readers to understand that a considerably flexible partial matching is possible by fixing its appropriate threshold value. As a more concrete illustration, let consider if an object in Fig.5<a>is identified as the OBAQ, who is a famouse charactor in the Japanese comic strip. As "HAND2" is occluded by something, 10 points will be reduced according to ① in Flg.5<c>, this operation corresponding to the case [1] stated above. Next with respect to *[2]* 1/4*10 points will be also cut down because a real relation is "CIN", which corresponds to ② Note that still more l/4*10*10/(10+8) points must be reduced from the total points by taking the effect into consideration which are brought about an evalutation of "MOUTH", where 1/4*10 points have been taken off with repect to ③ because of the mismatched relation "CIN". As a result of this evaluation the score for the "BODY" becomes 56.7 points. This means that the score is over the threshold value, then the object in Fig.5<a>can be identified with the "OBAQ".



```
(*SUBR  BODY  50)
        (*COL   WHITE   10)
        (HAND1  COUT   ((R)  C)  10)
        (HAND2  COUT   ((L)  C)  10) ①
        (HAIR   COUT   ((C)  U)   10)
        (EYE1   IN     ((R)  U)  10)
        (EYE2   IN     ((L)  U)  10)
        (MOUTH  IN     ((C)  C)   10)②
          .....
(*SUBR MOUTH 15)
        (*COL PINK  8)
        (MTH IN ((C) C) 10) ③
        .....
```

Fig.5 Object identification using models.

4.3.2 Object Identification

A process to identify objects assumed to be in the current scene is conducted by first extracting those object explicitly mentioned in the narrations. This indication by utterance means that their descriptions appear in the intermediate representations shown in 4.2. In advance to this execusion, it deserves to think of possibility if the plausible domain for objects can be restricted to a small portion of the image, because there are often cases where objects themselves or things related to them imply their existent environments. If such restrictive conditions are deducible, the object finding program can search them in this small region, in a top-down fashion. Otherwise, that is, no such condition is found, the program tries to go through a relative narrow space for those objects by assuming that only a little or not at all movements occur from the place they were located in the preceding images, or that animate objects continue moving toward the same direction as in the latest image. If all of those methods fail to find the specified object, and yet it is confirmed that they are in the given scene, the program works on its exhaustive blind search for it. These are the practical use of pictorial constraint implemented in this system, but those with respect to so-called common sense knowledge are, of course, now in effect. The following description for overcoats shows such an example:

```
OVERC:$$OUTER ´ `VEAR XMAN 10)
      $$IN-HOUSE   (IN XDRESSER 8)
                   (ON %CHAIR 5)
                   (WEAR %MAN -5)
```

A number put in the last position of lists represents the plausibility value for its occurence, where the most plausible one is given 10 points, whereas -10 for the least case as its score. As this description is self-explanatory, nothing might be needed. Along with these maneuvers, it is also important for the program to pay attention to the circumstance around the identified objects in order to find things closely related them, otherwise, the program cannot know a bit but just objects explicitly stated in the narrations, which is too poor. For an example, if some objects are found in the region seemingly corresponding to the sky and yet they cannot be there, the program must infer that they must be supported by something. For apples, which appear in our experiment shown later, something might be a tree or supporter like a table.

4.4 World model construction

Each time an image is given to the system, the image processing program records all things identifiable as an object in forms of their coordinate values with respect to a new coordinate system. On completion of these operations for successive frames, it is necessary for him to put those systems together into one global coordinate system as far as possible, in order to capture the global relations among things appearing in the given story. Since those images given along with narrations does not necessarily illustrate consecutive movement, a change in situation, like a case an actor enters a room from outside, often makes it difficult to tidy those systems up into such a unique system. In such a case, all frames following this situation are described with respect to this new system. The appreciation on the depth relations in the images is, in some sense, in effect by providing models with their standard dimensions as a clue to infer their approximate locations.

The procedure to implement the idea is like this: Now assume that the program sees a n-th frame, and its coordinate system is COn, we are working on it with respect to the standard system CO A.

(a) Find an object that appears in both COn and COm (m<n), and yet has a property "still" or "immovable" in its attribute slot. If such an object can be found, go to (b), otherwise (c).

(b) Calculate an origine of the COn with respect to the COA by referring to coordinates of this object in the COn, COm respectively. Then all object believed to be in this frame are put together with respect to COA.

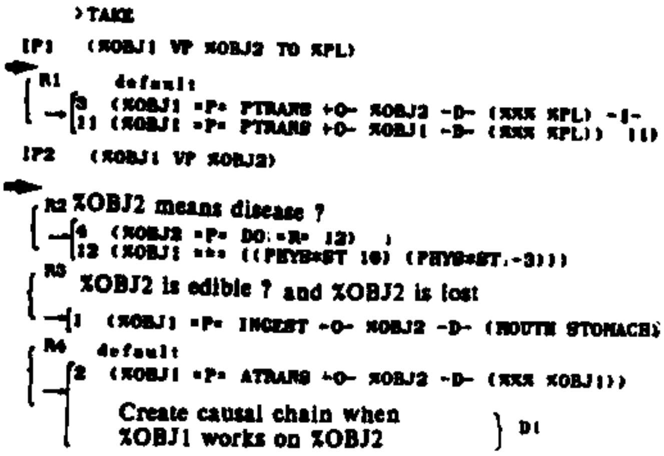
(c) Suspend the current operation and defer this until the position of some objects is determined with respect to COA.

(d) If there occurs a change in the situation, set a global system to this new one, say COA', and a rough relation between the old COA and this COA' is recorded in the COA', if possible.

4.5 Translation into The CD representation

We show a rough sketch of this translation process by referring to an example. See Fig.6, which is an easy version with the help of English that is equivalent to the content in 3.3.2. Now let assume that an intermediate expression (OBAQ TAKE APPLE3) is derived from the procedure given in 4.2. This expression leads the translation process to the execution of rules, R2, R3 and R4 in this order, because their key pattern IP2 successfully matches to this expression. As to the rule R2, it is clear that this is not applicable to the case at hand because a pattern variable XOBJ2, now bounded to APPLE3, does not mean disease. Then the next rule

R3 is checked for its legality. Because the value of %OBJ2 is edible, a result in the form of
(OBAQ =P= INGEST -O- APPLE3 -D- (MOUTH STOMACH))
is reached from the portion of the rule labeled with 1, if it is comfirmed that the %OBJ2 does not exist anywhere in the scene, which clearly means that it was eaten in this case. Note that the assumption on aspectual sense stated earlier in this paper is used here. The fact that the object is found to be as it was discourages the rule R3 from being attempted, and eventually rule R4 is activated. This is a so-called default rule, so that it concludes the following with no condition
(OBAQ =P= ATRANS -0- APPLE3 D- *(%%%* OBAQ))
On completion of execution of each rule, a demon, if attached to the rule, will be invoked in order to inspect causal relations among these CD descriptions asserted so far.



```
>TAKE
IP1   (%OBJ1 VP %OBJ2 TO %PL)
R1      default
   3  (%OBJ1 =P= PTRANS +O- %OBJ2 -D- (%%% %PL) -I-
   11 (%OBJ1 =P= PTRANS +O- %OBJ1 -D- (%%% %PL))   11)
IP2   (%OBJ1 VP %OBJ2)
R2 %OBJ2 means disease ?
   4  (%OBJ2 =P= DO: =R= 12)   ,
   12 (%OBJ1 =*= ((PHYS=ST 10) (PHYS=ST,-3)))
R3
   %OBJ2 is edible ? and %OBJ2 is lost
   1  (%OBJ1 =P= INGEST -O- %OBJ2 -D- (MOUTH STOMACH)
R4  default
   2  (%OBJ1 =P= ATRANS -O- %OBJ2 -D- (%%% %OBJ1))
      Create causal chain when
      %OBJ1 works on %OBJ2            } D1
```

Fig.6. Production rule for TAKE.

4.6 Answer Generation

By using of informations derived from events stated in the story and things closely related to them, this system can answer some questions concerning to this story.

4.6.1 Query generation

This is easily accomplished by first translating an input sentense into a query by substituting pattern variables for unknown portion of the sentence, as is the case with the interpretation of narrations.

4.6.2 Matching strategy

A matching procedure used to answer questions differs from the conventional matcher in the following four points.

(1) It is vital for him to correctly consider what portions of the story are interested in, because the plot proceeds along with time and actors can move around from place to place.

(2) As precise locative informations concerning to objects in the scenes are not asserted in the forms of CD notation as a rule, a routine that computes them directly from their coordinate values is required.

(3) With respect to descriptions stating the status of affairs, even if they were asserted in somewhere in the story once and for all, they are there as they have been until they will undergo some changes in their state in the consecutive frames. Our matcher must take care of this fact.

(4) Though assertions, in terms of the CD, first represent events that are stated in narrations or those observed in scenes, they are inserted additional relations among events when a reasoning process goes according to the plan that program should assert some facts with respect to causal relations or others. Consequently, when a matcher is going to find expected answers, it must match the query against a proper portion of descriptions by keeping itself free from such clutter.

S. Experimental result

In this section, beginning with the illustration on the detailed process by which representation is built up from the given pictures and narrations, ending with a dialogue referring to its world model constructed through reasoning.

5.1 World model construction

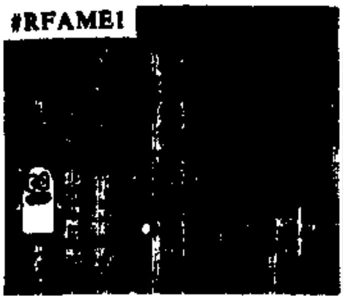
Fig.7 shows a series of pictures and narrations given to the system. As a result of the story being interpreted, the inner representation as shown in Fig.8, description concerning to Object Frame as in Fig.9 and the world model in regard to the location of Objects as in Fig. 10 are obtained. The figure shown in Fig. 10 shows a situation corresponding to first half of this story is stuffed into one global system. Note that the dimension of objects are drawn by considering their standard demensions in depth.

At the first frame, denoted as #FRAME1, program tries to find the OBAQ and APPLEs first comming out into this frame, whereas the ROCK, which isn't stated in words, is not regarded as to be inspected. The tree, however, is successfully recognized because program uses his common-sense knowledge that apples usually grow on trees. That is, in this case, the program first looks for a tree for the purpose of restricting the region the apples would locate, then find them by searching this tree region. From now on, when we refer to the descriptions given in Fig.8, 9, 10, we use the number preceding them.

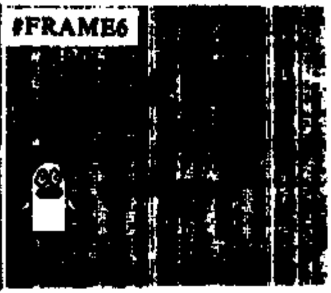
The floating point number in assertions 3,4,5 shows the position of APPLEs on the TREE1, and these apples are also recognized as a member of the object-group APPLE51, which corresponds to the word "apples" in the narration for the IFRAME1

At IFRAME2, the assertion 7 are derived through the rule which translate surface representation into the CD form, where "take" are successfully interpreted referring to the relation between STONE1 and his HAND. Concurrently with this operation, a demon is activated, shoving a fragment of description having the form of (=R= (12)) in the assertion 7. This is needed to record the fact that the action stated in 7 enables the event in assertion 12 to be caused.

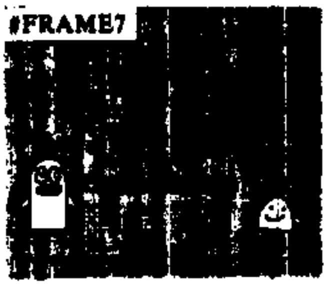
At #FRAME3, a transformational rule required to capture the word "throw" is used, and its consequence is correctly predicted and verified, and the representation 12 and a demon D2 are generated. This D2 convinces the program of the fact that the APPLE3 fell down on the ground in the intervals from #FRAME2 through #FRAME3. This also causes a new assertion 16, which states the APPLE3 was fallen with the STONE1, to be gone into the representaion. In addition it to these, creates an another demon which devotes its attention to the APPLE3 for conformation of what might happen to it.



OBAQ comes,He finds apples.

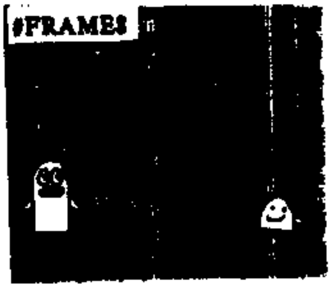


He puts it.

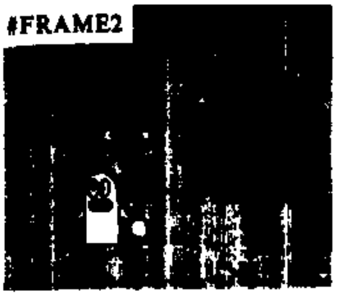


He calls OJIRO.

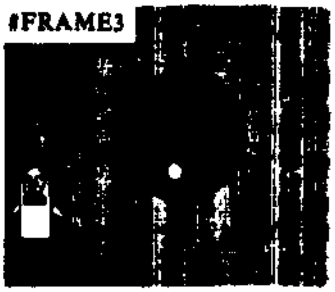
He gives OJIRO the apple.



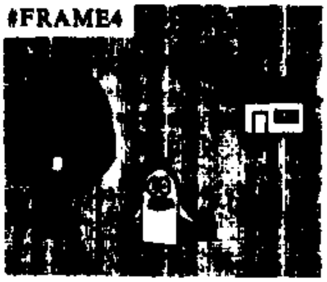
OJIRO takes it.



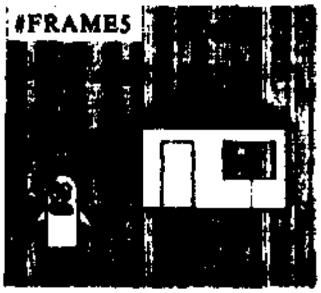
He takes a stone.



He throws the stone.



He takes an apple.



He takes it to his house.

Fig.7 Input scenes and narrations.

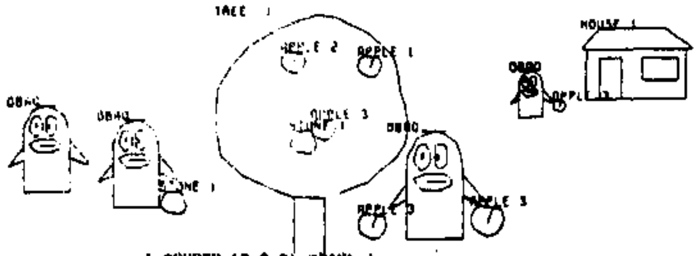
```
** REPRESENTATIONS **
 1) (OBAQ *P* PTRANS -O- OBAQ -D- (%PL %HERE))
 2) (OBAQ *P* MTRANS -O- APPLES1 -R- (%%% OBAQ))
 3) ((*POS) APPLE1 ON TREE1 +.729529E+00 +.264331E+00)
 4) ((*POS) APPLE2 ON TREE1 +.258823E+00 +.238853E+00)
 5) ((*POS) APPLE3 ON TREE1 +.435294E+00 +.563694E+00)
 6) ((*POS) APPLES1 ON TREE1)
 7) (OBAQ *P* ATRANS -O- STONE1 -D- (%%% OBAQ) *R* (12))

12) (OBAQ *P* SMOVE -O- STONE1 -D- (%%% %PL1) *E* (7) *R1* 16 *R* (18 23 26 31))
13) ((*POS) APPLE1 ON TREE1 +.715189E+00 +.271842E+00)
14) ((*POS) APPLE2 ON TREE1 +.263823E+00 +.245098E+00)
15) ((*POS) APPLES1 ON TREE1)
16) (OBAQ *P* PROPEL -O- APPLE3 -D- ((+.177382E+03 +.187341E+03 +.977839E+00)
(+.196673E+03 +.173271E+03 +.820665E+00)) -I- 12)
17) ((*POS) APPLE3 CONTACT OBAQ +.109574E+01 +.743750E+00)
18) (OBAQ *P* ATRANS -O- APPLE3 -D- (%%% OBAQ) *E* (12) *R* (23 26 31))
19) ((*POS) APPLE1 ON TREE1 +.743055E+00 +.265100E+00)
20) ((*POS) APPLE2 ON TREE1 +.312500E+00 +.244966E+00)
21) ((*POS) APPLES1 ON TREE1)
22) ((*PSS) HOUSE1 OBAQ)
23) (OBAQ *P* PTRANS -O- APPLE3 -D- (%%% HOUSE1) -I- 24 *E* (12 18))
24) (OBAQ *P* PTRANS -O- OBAQ -D- (%%% HOUSE1))
25) (OBAQ *P* PTRANS -O- OBAQ -D- (%%% (INSIDE HOUSE1)))
26) (OBAQ *P* PTRANS -O- APPLE3 -D- (%%% %PL) *E* (12 18))
27) ((*POS) APPLE3 ON TABLE1 +.456522E+00 -.215909E+00)
28) (OJIRO *P* PTRANS -O- OJIRO -D- (%%% OBAQ) -I- 29)
29) (OBAQ *P* PROPEL -O- VOICE -D- (OBAQ OJIRO))
30) ((*POS) APPLE3 ON TABLE1)
31) (OBAQ *P* ATRANS -O- APPLE3 -R- (OBAQ OJIRO) *E* (12 18))
32) (OJIRO *P* INGEST -O- APPLE3 -D- (MOUTH STOMACH))
```

Fig.8 A portion of primary pass.

```
*FR1
 > OBAQ          ** (*ACT (1 2) *PLIST ((BODY . 9) (HAND1 . 11) (HAND2 . 10)
  (HAIR . BR7) (((H3 H2 H1) 32 31 30)) (EYE1 . 12) (BEYE1 . 13) (EYE2 . 14)
  (BEYE2 . 15) (MOUTH . 16) (MTH . BR8) (((MTH1) 34))) *WHERE (20 120)
  *SIZE (41 74) *3D (+.205142E+02 +.131291E+03 +.102571E+01))
 > APPLES1       ** (*ACT (2) *MEM (APPLE1 APPLE2 APPLE3) *POS (6))
 > APPLE1        ** (*GNAM APPLES1 *POS (3) *PLIST ((APP . 3) (HT1 . BR2) (((A1)
  8)) (HT2 . BR1) (((A2) 7))) *WHERE (207 92) *SIZE (13 21)
  *3D (+.203967E+03 +.906820E+02 +.980348E+00))
 > APPLE2        ** (*GNAM APPLES1 *POS (4) *PLIST ((APP . 4) (HT1 . BR4) (((A1)
  10)) (HT2 . BR3) (((A2) 9))) *WHERE (166 87) *SIZE (12 20)
  *3D (+.164048E+03 +.851086E+02 +.978261E+00))
 > APPLE3        ** (*GNAM APPLES1 *POS (5) *PLIST ((APP . 5) (HT1 . BR6) (((A1)
  12)) (HT2 . BR5) (((A2) 11))) *WHERE (180 138) *SIZE (12 20)
  *3D (+.179022E+03 +.135000E+03 +.978261E+00))
 > TREE1         ** (*POS (3 4 5 6) *PLIST ((LEAF . 2) (TRUNK . 1))
  *WHERE (152 61) *SIZE (85 157) *3D (+.154089E+03 +.538121E+02 +.882166E+00))
```

Fig.9 A portion of OF.



```
1 *COORD (0 0 0) *BASE 1
2 *COORD (+.034375E+01 +.272610E+01 +.169016E-01) *BASE 1
3 *COORD (+.429359E+02 -.470614E+01 -.479455E-01) *BASE 1
4 *COORD (+.111251E+03 -.973746E+01 -.110794E+00) *BASE 1
5 *COORD (+.550331E+03 +.554462E+02 +.578296E+00) *BASE 1
6 *BASE 6 *COORD (0 0 0) *DVLP (INSIDE HOUSE1)
7 *COORD (+.000000E+00 +.000000E+00 +.000000E+00) *BASE 6
8 *COORD (+.000000E+00 +.000000E+00 +.000000E+00) *BASE 6
```

Fig.10 The positional world model.

At #FRAME5, there is nothing unmovable except for HOUSE1 in both #FRAME4 and #FRAME5, but as this HOUSE1 is recognized for the first time at #FRAMES, it is used to relate #FRAME5 to the global coordinate system denoted as *BASE in Fig. 10. Here the reader must note that "the house" drawn in #FRAME4 is not perceived to be there as it is not mentioned in the narration at #FRAME4. Consequently it is not found in #FRAME4 untill program look at things corresponding to HOUSE1 at #FRAME4 with utmost care, because HOUSE1 is the only object believed to be immovable at #FRAME5.

At #FRAME6, a change in the situation occurs and a new basic coordinate system is generated, in this case no relation between these basic systems is recorded (see 7-th list in Fig. 10).

Here the TABLE1 in 27-th assertion was not mentioned in the narration, but is found to be in a scene, because its existence is predicted and verified through the location of APPLE3 and the fact that the action "put" suggests a place on which the object is placed.

5.2 QA result

This program can answer some questions shown in Fig.11, where a first few examples are listed along with the reason why the answers can be reached by the program. As with the section 5.1, we refer to examples with their number.

(1) The question beginning with "how many" can be answered by counting how many assertions or descriptions match to the query. As to this, the context is recognized as #FRAME4. using "When OBAQ took the apple", and this context enables the matcher to find a query ((•POS) ?-APPLE ON TREE1) to be the same as the assertions 19. 20.

(2) By finding that there is a description =E= in the 18-th question, which indicates that the action stated in 12 causes the event in 18 to come about, it is known that the expression 12 gives the answer.

(3) As with the above example, assertions 23, 26 and 31 are used as an answer simply by finding a descriptor =R= representating causal relations. In this version, how much detail should be given in explanation is not considered very much.

(4) The 32-th assertion easily leads to this answer. It is easy for a program using the CD representation to match "EAT" to "TAKE", thus we believe that the CD concept fits well the sophisticated knowledge representation requiring many sorts of knowledge sources like our work.

(5) A frame number is used to answer question. The level of an English generation routine it uses is not so strong yet that some answers are often given in this way.

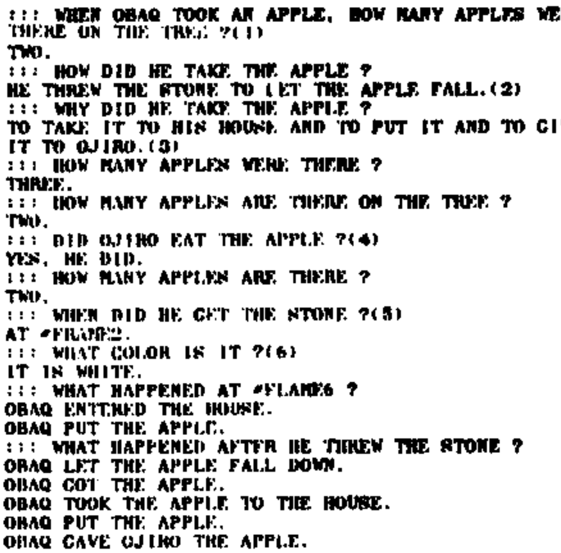
(6) There being no assertion concerning to the color of STONE1, this answer is derived directly through seeing a line drawing, itself.

8. Conclusion

The reference on both image and language enables the following effect to be realized.

(1) An explicit mention of objects can simplify the objects identification process by first restricting the region they will exist to the small one, and by next infering things related to them.

(2) The confinement of description of objects into the forms of coordinates contributes to a compact representation of events and a combinatorial explosion can be avoided that will occur when all relations among things in scene are asserted in words in the data base.



```
::: WHEN OBAQ TOOK AN APPLE, HOW MANY APPLES WERE
THERE ON THE TREE ?(1)
TWO.
::: HOW DID HE TAKE THE APPLE ?
HE THREW THE STONE TO LET THE APPLE FALL.(2)
::: WHY DID HE TAKE THE APPLE ?
TO TAKE IT TO HIS HOUSE AND TO PUT IT AND TO GIVE
IT TO OJIRO.(3)
::: HOW MANY APPLES WERE THERE ?
THREE.
::: HOW MANY APPLES ARE THERE ON THE TREE ?
TWO.
::: DID OJIRO EAT THE APPLE ?(4)
YES, HE DID.
::: HOW MANY APPLES ARE THERE ?
TWO.
::: WHEN DID HE GET THE STONE ?(5)
AT #FRAME2.
::: WHAT COLOR IS IT ?(6)
IT IS WHITE.
::: WHAT HAPPENED AT #FLAME6 ?
OBAQ ENTERED THE HOUSE.
OBAQ PUT THE APPLE.
::: WHAT HAPPENED AFTER HE THREW THE STONE ?
OBAQ LET THE APPLE FALL DOWN.
OBAQ GOT THE APPLE.
OBAQ TOOK THE APPLE TO THE HOUSE.
OBAQ PUT THE APPLE.
OBAQ GAVE OJIRO THE APPLE.
```

Fig.11 QA result.

(3) The relation found in the given scene can help the interpreter of language resolve interrelation among prepositional groups.

(4) The usage of production-like representation for the translation from English into the *CD* description and avaiiability of pictorial information in this translation makes these operations quite straightforward.

On concerning to (l).the following must be noted: from the assumption stated first in this paper, there is a case where some objects are not recognized what they are, when they are not stated explicitly and yet not deducible from any facts known so far, even if they are identifiable when compared with their models. This point should be regareded as a good point rather than a defect, because the system at any time can answer questions about such things as the clock in #FRAME6-8. This clock was explicitly recognized at any time in the story, but the following question such as "Is there anything on the wail?" or "Is there a clock anywhere?" can be answered by ulitizing the words appearing in these questions as a clue to find and recognize objects. But there is obviously great deal more to be improved in order to completely acheive the goals of this research.

References

(1)R.Schank: Conceptual Information Processing, North-Holland Publishing Co.Ltd. 1975

(2)W.Lehnert: Human and Compttational Question Answering, Congnitive Science 1, 47-74(1977)

(3)E.Charniak: Toward a Model of Children's Story Comprehension, AITR-2666, MIT, (1972)

(4)J.Carboneil: Towards a Process Model of Human Personality Traits, Artificial Intelligence 15, 49-74 (1980)

(5)R.Schank: Controling Inference, Artificail Intelligence 12, 273-297 (1979)

(6)R.Wilensky: Why Jhon Married Mary: Understanding Stories Involving Recurring Goals, Cognitive Science 2, 235-267 (1978)

(7)D.Waltz, L.Boggess: Visual Analog Representations for Natural Language Understanding, 6th-IJCAI, 926-934(1979)

(8)S.Tsuji,S.Kuroda and A.Morizono: A Simple Cartoon Film Understanding System, 5th-IJCAI, 609-610 (1977)

(9)S.Tsuji,M.Osada and M.Yachida: Three Dimensional Movement Analysis of Dynamic Images, 6th-IJCAI, 896-901 (1979)