Constructing a lexicon of action

Tokunaga Takenobu, Okumura Manabu, Saitô Suguru, Tanaka Hozumi

Tokyo Institute of Technology
Tokyo Meguro Ôokayama 2-12-1, 152-8552, Japan
{take@cl.cs.titech.ac.jp}

Abstract

This paper describes a Japanese speech dialogue system that enables a user to interact with agents in a virtual world and proposes a design framework for building a lexicon of action. This lexicon is used to realize the behavior of the agents in response to the user's commands. The lexicon has two levels – a macro and micro level. The system uses the macro-level lexicon, which is similar to a conventional plan library, to translate the user's goal to a sequence of basic movements. This process is the same as conventional planning with symbol manipulation. The micro-level lexicon is used to translate the basic movements into animation, which is represented by a sequence of avatar postures. We discuss how to define a set of basic movements and how to make these basic movements reusable.

1. Introduction

The question "What is language understanding?" is difficult to answer. We claim that the meaning of a language expression can be explained in terms of an agent's actions on responding to the expression. To substantiate this idea, we are developing a virtual actor system in which a user can interact with agents in a virtual world (Figure 1) (Shinyama et al., 2000). Through speech input, a user can command the agents to manipulate objects in the virtual world. The agent's behavior and the subsequent changes in the world are presented to the user through a camera in terms of a three-dimensional animation, which is considered as the result of understanding the user's utterances.

This system shares many ideas with Winograd's SHRDLU (Winograd, 1972) in which different types of blocks can be manipulated by a robot arm on the basis of a user's keyboard input. The emphasis in this system is on treating various kinds of linguistic phenomena, such as anaphora resolution and discourse analysis. The system also shows the behavior of the robot arm on a graphic display, but the behavior was very simple and deterministic.

Since SHRDLU translates the user's utterances into procedures to manipulate blocks in a straightforward way, it does not require a lexicon to describe the behavior of the robot arm. Similarly, in our current prototype, shown in Figure 1, the agents have a simple physical structure and their actions are limited. Therefore, we do not use any lexicon to describe the agents' behavior.

Unlike SHRDLU, however, we are aiming to produce more complex agent behavior as a result of language understanding. To achieve this, we need to extend the current system by establishing a set of principles for building a knowledge to translate a linguistic expression into an animation. In this paper, we propose a design framework for building a lexicon for translating a linguistic expression into a three-dimensional animation.

In the following sections, we first give an overview of our system. Section 3. describes the issues that must be resolved and our approach to them. We conclude with a brief summary of future research direction.

Figure 1: A screen shot of a prototype system

2. System architecture

Figure 2 illustrates an overview of the system. The speech recognition module receives the user's speech input and gives a sequence of words. The language understanding module analyzes the word sequence and extracts the user's goal (intention). This goal could itself be considered as the meaning of the utterance, but we go a step further and realize it as an animation.

To achieve this, the planning module builds a plan to generate an animation by referring to the lexicon describing actions. In other words, the planning module translates the user's goal into an animation. However, the properties of these two ends are very different and straightforward translation is rather difficult. The user's goal is represented in term of a symbol or a structure of symbols, while the animation data is a sequence of numeric values. To bridge this gap, we take a two-stage approach – macro and micro planning. The lexicon is also divided into two classes, a macro and a micro level, corresponding to the planners.

The macro-planning module translates the user's goal extracted by the language understanding module into a sequence of basic movements. This process is the same as conventional planning (Fikes, 1971), i.e., on being given a goal, giving a sequence of predefined primitive operators. In this case, the basic movements correspond to primitive operators. For example, the goal "hold(cup)" would be sat-

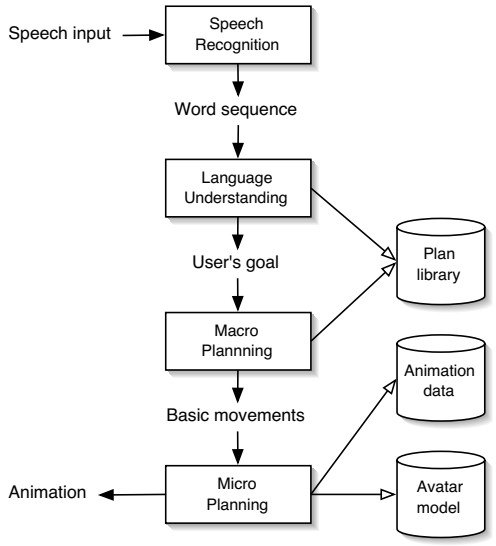

Figure 2: A system overview

isfied by an action “pick up a cup in the cupboard”, and this action could be decomposed into “go to the cupboard”, “open the cupboard”, and “grasp a cup”, which could be basic movements. Therefore, a macro-level lexicon is similar to an ordinary plan library.

The micro-planning module translates a basic movement into animation data, which provide time-sequence information for the avatar’s joints. We have adopted the format of a human model named “H-Anim (Figure 3)” (Humanoid Animation Working Group, 2001) which models a human body with about 100 nodes. Each node corresponds to a bone between joints that has a certain amount of freedom. Setting the angle for each joint defines the avatar’s posture. Thus, a sequence of joint angles defines an animation. The lexicon referred to by the micro planner needs to include time-sequence data for the angles of each joint. Because it is very time consuming to create such motion data from scratch, we used a motion capture system to collect motion data. This involved placing several sensors on a human body and gathering motion data through the sensors.

There are some difficulties in distinguishing the level of the two planners. What should be handled at the macro level and what at the micro level? The division of tasks between these planners is based on whether the task involves coordinate values for the virtual world. The macro planner deals only with symbols and mapping from symbols into coordinate values is handled by the micro planner. For example, relations of locations are handled in terms of symbolic relations such as “right_of”, “in_front_of” in the macro planner. The micro planner then calculates coordinate values from these symbolic relations (Yasima, 2002).

3. Issues to be resolved and our approach

To establish this framework we must first resolve the following issues:

(1) How to define basic movements?

(2) How to make basic movements reusable?

In this section, we describe our approach to these issues.

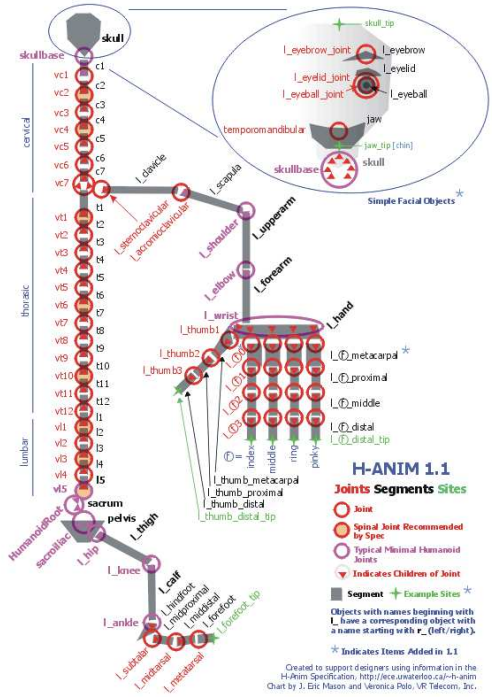

Figure 3: Human model format “H-Anim”

3.1. Basic movements

In a conventional planning framework, a set of primitive operators is defined *a priori*. However, it is not an easy task to define a set of basic movements when generating an animation based on planning. This issue has been discussed for years and is still controversial among philosophers (Israel et al., 1993). For example, the movement “open the cupboard” could be further decomposed into “stretch out the arm”, “grasp the knob of the cupboard door” and so on.

To define basic movements, we took both a top-down and a bottom-up approach. In the top-down approach, we analyzed a Japanese basic verb lexicon named IPAL (Hasimoto et al., 1996), which includes 861 basic verbs and describes various features of verbs, such as subcategorization and aspectual information. According to these features, the verbs have been further divided into 3,379 subentries. In addition to these basic verbs, the IPAL supplies a few deverbal noun entries. There are 94 subentries of deverbal nouns. We analyzed these 3,473 subentries.

We first filtered out the subentries on the basis of whether a verb takes a noun with a semantic marker “+human” as its subjective case. This was done automatically by referring to the subcategorization information. After this filtering, 2,437 subentries remained. These were further checked manually. As a result, 1,291 subentries remained as candidate verbs to suggest basic movements. In the manual inspection, the following features of the verbs were referred to:

- orthographic form (in Kanji)
- semantic description in natural language
- semantic category (“movement”, “mental state”, etc.)

- semantic class of thesaurus
- example sentences

The set of basic movements depends on the domain that the system deals with. In taking a bottom-up approach, we assumed a scenario in which two persons interact in a kitchen for a couple of minutes, and enumerated the verbs used to describe the scene. Figure 4 shows a fragment of the continuity for this scenario.

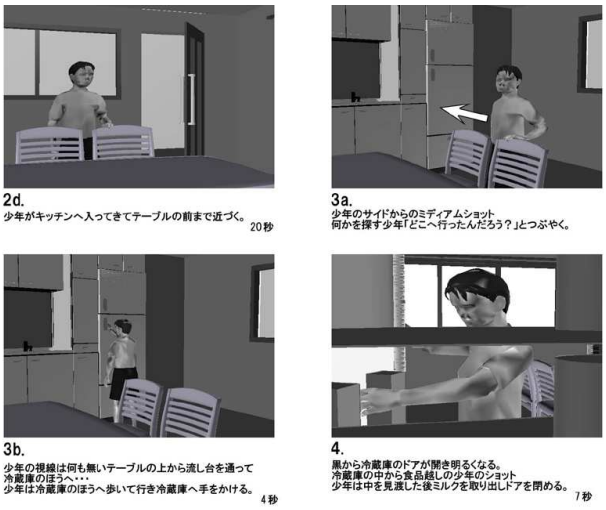

Figure 4: A fragment of the continuity for a scenario

Candidates for the basic movements were extracted from this animation sequence using the following procedure:

1. Identify expressions specifying the action of the characters
2. Check that the expression describes a physical action
3. If it describes a physical action, it is identified as a candidate for the basic movement. Otherwise, the action is decomposed into the physical actions required to achieve the action.

Deciding if a verb describes a physical action is also a difficult task. In a sense, it is the same as defining basic movements. As mentioned in section 2., the distinction between macro and micro planning is based on whether absolute coordinate values are involved. We used the same criterion to judge a physical action. This judgment was made subjectively.

For example, "open a door" could be further decomposed into "grab a door knob" and "push (pull) the knob". However, we also need to take into account the coordinate values and absolute values of the joint angles when realizing "grabbing a door knob".

Table 1 shows a list of candidates extracted from observing the kitchen scenario. The hyphens in the candidates denote the boundaries of the morphemes. The second column labeled "IPAL" denotes whether the candidate has also been extracted by the top-down method. As the table shows, there is quite a lot of overlap between the two lists of candidates.

Candidate	IPAL
idou-suru (move)	no
aruku (walk)	yes
iku (go)	yes
arukimawaru (wander)	no
tikazuku (approach)	yes
tikazukeru (make something closer)	no
muku (turn, direct)	yes
mukeru (turn, direct)	yes
iu (say, utter)	yes
tukamu (grasp)	yes
osu (push)	yes
furu (shake)	yes
oku (put)	yes
tukidasu (stick out)	no
hanasu (release)	yes
hiku (pull, draw)	yes
motiageru (raise, lift)	no
te-wo-mageru (bend one's arm)	no
te-wo-nobasu (stretch one's arm)	no
kubi-wo-mawasu (turn one's face)	no
nodo-wo-ugokasu (swallow down)	no
kao-wo-kowabaraseru (frown)	no
kuti-wo-akeru (open one's mouth)	no
kuti-wo-toziru (close one's mouth)	no

Table 1: Candidates for basic movements

The first group includes verbs related to transfer. There are three verbs that are not included in the candidate list from IPAL. "*Idou-suru* (move)" has a construction "a deverbal noun *idou* (move) + a light verb *suru* (do)". Since there are many deverbal nouns in Japanese, the IPAL includes only a few of them. "*Arukimawaru* (wander)" is a compound verb, "*aruku* (walk) + *mawaru* (around)". This compound verb itself is not included in the IPAL list, but both element verbs are included. "*Tikazukeru* (bring something closer)" is a causative form of "*tikazuku* (approach)" which is included in the IPAL list.

The second group includes transitive verbs. The two verbs not included in the IPAL list are compound verbs. The elements of these verbs are included in the IPAL list.

The candidates in the third group have the construction "a noun representing a part of the body + a case marker *wo* (marking objective case) + a verb". The verbs in all of these candidates except for "*kao-wo-kowabaraseru* (frown)" are included in the IPAL list. Since we are considering human movements, it is reasonable to treat these candidates as basic movements.

Another issue related to basic movements is the problem of vagueness. In past research on natural language processing, vagueness has not attracted much attention compared with ambiguity. When attempting to realize a particular behavior as an animation, we need to narrow the interpretation sufficiently to realize an animation.

For example, suppose we have admitted "grasp" as a basic movement, the actual movement of every joint of a body can be realized in innumerable ways. In particular, the actual movement depends very much on the object to be grasped. To avoid this problem, we approximated ob-

jects as simple geometric objects, such as a column, a ring, a sphere, etc. The difference between such abstracted objects and the actual objects is filled up by introducing active interpolation of avatar's postures.

To explore the variations of a movement, we conducted a preliminary experiment in which we asked two subjects to grasp a cup in as many different ways as possible and took pictures of the movements. We collected 40 variations and classified them into 10 classes as shown in Figure 5. Ideally speaking, this classification should be fully automated, but at present we have no idea how to automate the process. Some machine learning techniques may be applicable. This is the subject of future work.

Figure 5: Classification of "grasp"

3.2. Reusability of basic movements

The second issue is the reusability of basic movements. As mentioned before, one of the advantages of a motion capture system is that it enables us to collect motion data easily. However, it is quite difficult to modify the captured motion data and to combine multiple motion data to depict simultaneous movements. Since an action is decomposed into basic movements, the basic movements need to be combined to realize a complex action. Thus, the compositionality of basic movements is indispensable to our framework.

To achieve compositionality, each motion data is annotated with features corresponding to the avatar's joints, and the precedence between the motions for a feature is defined. For example, suppose we have motion data for the movements "walk", "run" and "wave". An action "running while walking" is impossible but "waving while walking" is possible. This could be explained as follows. "Walking" and "running" movements are performed using the same features (feet, legs, etc.) and using these features is essential for the movements. Therefore, conflicts in these features prevent the realization of simultaneous movements. "Waving", on the other hand, mainly uses different features (arms, hands, etc.). The use of some features, such as arms, might conflict with "walking", because a human usually moves the arms while walking. However, using the arms is not essential for "walking", while it is essential for "waving". Thus, the "waving" movement takes priority over the "walking" movement in features corresponding to arms.

Another aspect of the reusability of basic movements relates to variations of a movement. For example, "sitting on a chair" and "sitting on the floor" would be depicted as different behavior. However, there is a overlap of joints used in performing these behaviors. Thus, it is possible to use the same animation data to generate these different behaviors. If we could abstract the difference between these behaviors from the animation data in terms of the avatar's joint angles, the abstract movement "sit" would be used in both "sitting on a chair" and "sitting on the floor". To achieve this, we need to study the captured data in more specific detail.

4. Concluding remarks

This paper described the process of building a lexicon of action to be used in a speech dialogue system with visualization of the virtual world. Unlike the lexicon for text processing, which has been investigated by many researchers, a lexicon of action must bridge the gap between symbols and numeric time-sequence data. To achieve this, we devised a system consisting of two planning modules and introduced basic movements as an interface between them. Then, we discussed the definition of basic movements and their reusability.

The project is still ongoing and several aspects of the research require further study, as mentioned above. In particular, we need to automate the classification of variations of a basic movement, and their abstraction. Machine learning techniques may be useful for this purpose. However, it is not yet clear what kind of information we should collect and what kind of features are most effective. We need to manually investigate collected animation data to shed light on these issues before moving on to the automation phase.

5. References

R. E. Fikes. 1971. STRIPS: A new approach to the application of theorem problem solving. *Artificial Intelligence*, 2:189–208.

M. Hasimoto, Kuwahata, K. W. Murata, F. Aoyama, and T. Tonoike. 1996. Some remarks on ways to compile Japanese lexicons for computers. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 115–122. http://cactus.aist-nara.ac.jp/lab/events/SNLR/snlr.html.

Humanoid Animation Working Group. 2001. H-anim. http://www.h-anim.org.

D. Israel, J. Perry, and S. Tutiya. 1993. Executions, motivations, and accomplishments. *The Philosophical Review*, 102(4):515–540, October.

Y. Shinyama, T. Tokunaga, and H. Tanaka. 2000. Kairai – Software robots understanding natural language. In *Third International Workshop on Human-Computer Conversation*, pages 158–163.

T. Winograd. 1972. *Understanding Natural Language*. Academic Press.

E. Yasima. 2002. Expression of relative location in virtual world. Master's thesis, Tokyo Institute of Technology. (in Japanese).