

Animated Agents that Understand Natural Language and Perform Actions*

†Hozumi Tanaka, †Takenobu Tokunaga and ‡Shinyama Yusuke

†Department of Computer Science, Tokyo Institute of Technology

‡New York University

{tanaka, take}@cl.cs.titech.ac.jp, yusuke@cs.nyu.edu

Abstract

The natural language understanding (NLU) research environment has changed drastically in the past two decades. Better technologies in speech recognition, natural language processing and computer graphics are now available. We are in a good position to begin developing a lifelike animated agent who performs actions specified by natural language commands given by a speaker. It combines three technologies: NLU, speech recognition and computer graphics. A lifelike animated agent system named Kairai was developed at our laboratory to carry out preliminary research on the next generation NLU system. After giving a brief introduction of the Kairai system, we will conclude by outlining what problems ought to be solved in the future.

1 Introduction

Historically, the most important NLU system was SHRDLU developed by Winograd at MIT in the early 1970's (Winograd, 1972). This system was a kind of a software robot that worked in a toy block world simulated in a virtual space. However, rather than head, feet and hands, the robot was equipped with a small stick. SHRDLU was not regarded as a lifelike animated agent, but it has distinctive features. It could understand English dialogue input from a user's terminal (no speech input) according to which it performed very simple tasks such as "Pick up a red block on the table" and "Put it in the green box". The system could also answer simple queries about the current state of the toy block world. It enabled to resolve anaphoric

ambiguities, and build a plan to carry out a task specified by the input sentence. SHRDLU demonstrated the promising future of NLU research.

The NLU research environment has changed drastically in the past two decades. Better technologies in speech recognition, natural language processing and computer graphics are now available. We have obtained a huge amount of computing power under which research on computer graphics has made significant progresses in generating complex and realistic 3-D animated robots in a virtual space.

We review two typical related works (Badler et al., 1993)(Badler et al., 1999). Badler et al. were going to build 3-D animated agents who understood natural language and could perform some actions in a virtual space. The agent was given language-base instructions from which they extracted parameters for its actions. The parameters contains many information such as linguistic information, spatio-temporal information, manner information that was often expressed as adverbs, and both applicability and terminating conditions. Even though it is very important to handle ellipsis and anaphoric expressions, which very often occur in spoken language or in conversations, Badler et al. do not pay any attention to those expressions(Badler, 1998).

Cassell et al. (Cassel et al., 1999) pointed out that conversational skills that human has play very important in human-computer interactions. Such skills include the abilities to use face, hands and melody of the voice to regulate the process of conversation, as well as the ability to use verbal and non verbal means. They developed a system called "Rea", which was an embodied conversational agent and implemented actually social, linguistic, and psycho-

* This is a Grant-in-Aid for Creative Scientific Research supported by th Ministry of Education, Culture, Sports, Science and Technology, and Japan Society for the Promotion of Science

logical conventions of conversations. Rea with a human-like body can respond to human by using eye gaze, body posture, hand gestures and facial expressions. While Rea emphasizes the importance of non-verbal functions in conversations, the system does not mention the problem of vagueness in agent's actions. To visualize their actions we have to solve the problem.

After providing readers with a brief overview of Next generation NLU in 2 and our 'Kairai in 3 along with its sample dialogues (4), some problems in Kairai will be discussed in 5. Kairai enables to make proper interpretations for such adverbs as "left" and "right". Cameraman is invisible but manipulates his camera to take the view of the current virtual space. In 6, it is discussed that one-to-many conversational pattern is important and should be considered more in the future.

2 Next Generation NLU System

The question that needs to be answered is: What kind of research is necessary to build an intelligent software robot that understands natural language expressions. Consider the following dialogues:

1. Human: "Open the curtain covering the window a little."
>Robot goes to the curtain, and grasps it by his/her hand and opens it.
2. Human: "A little bit more."
>Robot opens the curtain a little bit more.
3. Human: "Too much."
>Robot closes the curtain a little.
4. Human: "Air in the room is polluted."
>Robot opens the window.

The first command issued by the Human makes the Robot create a plan to go to the curtain, grasp and open it. Such a plan is called a macro level plan. There are many ways to grasp and open the curtain. Robot has to select one of them to generate a micro level plan in order to carry out his/her actions. We can conclude that the Robot certainly understand natural language by watching the Robot's actions corresponding to what Human says. In other words, the Robot's actions, which are visualized in a virtual space as an animation, ver-

ify the NLU ability of the Robot. The visualized actions provide us with a NLU evaluation method more severe than that of Turing test, since the latter does not take account of visualized behavior of AI systems(Badler et al., 1993).

The second and third commands lack a verb in addition to a subject and an object. Robot has to augment these elliptical words by considering the context of a dialogue. With respect to the second command, Robot has to infer "open" as an appropriate elliptical verb, and then carries out the action "open." On the other hand, in the third command, "open" is also a correct elliptical verb, but Robot has to perform an opposite action "close" in this case. In other words, the Robot has to extract the intended actions from indirect speech act commands (Cohen et al., 1990). The second and third commands are related to the problem of vagueness, which were sometimes overlooked in the past NLU research.

The fourth command includes a typical indirect speech act that should be understood correctly for Robot to perform "open the door." Extracting true intentions in indirect speech act is one of the very difficult computational tasks.

In summary, the next generation NLU systems has to consider at least following items.

1. Resolution of anaphoric expressions by using the theory developed by Grosz (Grosz, 1986) along with the centering theory.
2. Augmentation of ellipsis.
3. Extracting true intentions from a command. The above three items are issues in situation dependent natural language processing (SDNLP), since they have to consider the context of dialogue along with the current state of virtual space (environment) where robots exist. SDNLP is a key technology for future NLU systems.
4. Combining technologies of speech recognition, NLU and computer graphics.
5. Handling ill-formed sentences that include fillers, additions, repairs and repetitions.

The fourth item brings about the fifth item.

3 Kairai System

For the feasibility study on the next generation NLU system tightly combining three technolo-

gies: speech recognition, NLU and computer graphics, we developed a prototype NLU system called Kairai (Shinyama et al., 2000)(Shinyama et al., 2001)(Tanaka, 2000)(Tanaka, 2001). Kairai system incorporates several 3-D software robots with which we can converse. It accepts voice inputs (spoken inputs), interprets them and performs the tasks specified in the virtual space.

The task executions are visible on a display screen as an animation. There are four software robots in Kairai system. In addition to three visible software robots: a horse, a chicken, and a snowman, a cameraman is also a software robot controlling his camera to give a different perspective of the virtual space. The cameraman and his camera are invisible on the display screen. The camera handling is specified through commands such as “Go near the horse.” In consequence, the figure of the horse is enlarged.

Kairai understands what we say in natural language, especially the words such as “left”, “right”, “in front of” and “behind” that indicate relative location in a virtual space. Typical actions performed by the (visible) software robots are “Push”, “Go”, and “Turn.” Interesting thing is that interpretations of “left” and “right” are determined by considering both the orientation of a software robot and human who issued each command.

Figure 1 shows the outline of Kairai system whose architecture is not new and is divided into three parts: speech recognition module, NLU module, and animation generation module. The speech recognition module transforms speech input into a sequence of words that become input to the NLU module that analyzes the input by using both a grammar and a dictionary and then extracts a meaning structure called a frame structure along with anaphora resolution and ellipsis handling. The latter two are together called a discourse process and refer to the context of past dialogues between human and Kairai. After a task plan is created by the NLU module, it is forwarded to the animation generation module to yield an animation on the display.

Figure 2 is a snapshot of animation generated by Kairai system. Readers can see three software robots in the virtual space. According

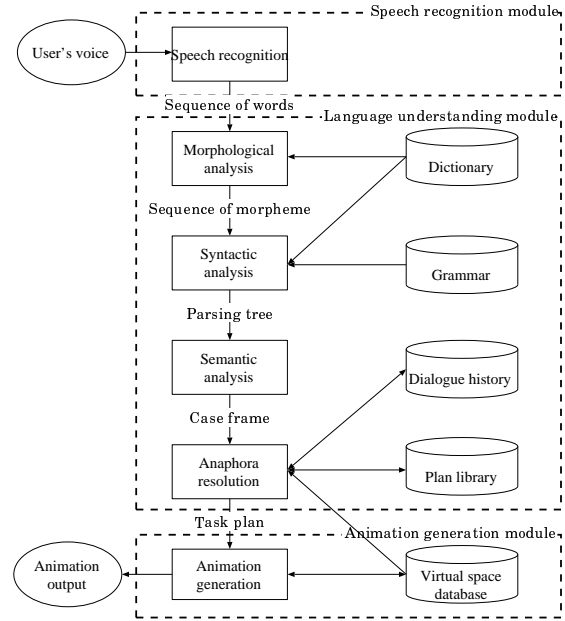


Figure 1: Outline of Kairai System



Figure 2: A Snapshot of Kairai

to the commands provided by the human, software robots including a cameraman move and perform appropriate actions in common space.

4 Sample Dialogue with Kairai

The typical dialogue that Kairai understand is shown below.

In the virtual space, there are four software robots (animated agents), Horse, Chicken, Snowman and Cameraman, the last of which is invisible but manipulates his camera to take the view of the current virtual space. In addition to them, there are two red spheres and two blue spheres on the ground in the space. Through voice input provided by Human, Kairai accepts an imperative sentence one by one.

1. Human: Horse, push the sphere located in the left to the front of Chicken.

>Kairai decides which sphere is specified by the command and to where Horse push it by making reference to the current state of the virtual space. According to the command interpretation by Kairai, he/she makes the Horse do the push action. Suppose the color of the sphere is blue.

2. Human: Push the red sphere, too.

>Considering Horse’s view, Kairai decides the red sphere that Horse should push and the Horse does the action. Note that “the red sphere” is an example of deictic expressions.

3. Human: Chicken, push it, too.

>Kairai resolves the anaphoric ambiguity given by “it” using a context, namely the preceding commands. In this case, “it” indicates the red sphere, which Horse pushes. Chicken does the action.

4. Human: Further.

>Although there is no subject, no object and no verb, Kairai augments these elliptical words by considering the context accumulated through the dialogue between Human and Kairai. Kairai forces the Chicken to push the red sphere further. Due to the visualization, Kairai also determines how far the Chicken pushes the sphere. This problem is called “vagueness” by linguists. Thus, Kairai carries out anaphora resolution, ellipsis augmentation, and vagueness handling.

5. Human: Cameraman, move close to the red sphere.

>Kairai makes the camera move close to the red sphere. As the result, it zooms in the red sphere and changes the view of the virtual space.

5 Problems in Kairai System

The experiences with respect to Kairai system which was developed as a prototype system to study the feasibility of the next generation NLU systems, brought realization of many remaining problems. The first and the most important

problem is that Kairai was not really a multi-agent system composed of autonomous agents (Febler, 1999)(Weiss, 1999).

As each software robot (agent) in Kairai seems to carry out his action independently, Kairai system, at a glance, is a multi-agent system. However, Kairai system is not an actual multi-agent system. In addition to four software robots mentioned before, there is another special agent who knows everything in the virtual space, receives and processes a sequence of words sent by the speech recognition module. This is illustrated in Figure 1.

After accomplishing NLU tasks, the special agent decides which software robot should perform what kind of actions and then activates an appropriate software robot. As any software robot in Kairai system executes the task plan generated by the special agent, it is not really an autonomous agent in a virtual space. This is the reason why Kairai is not really a multi-agent system. The problem discussed here brings about another problem.

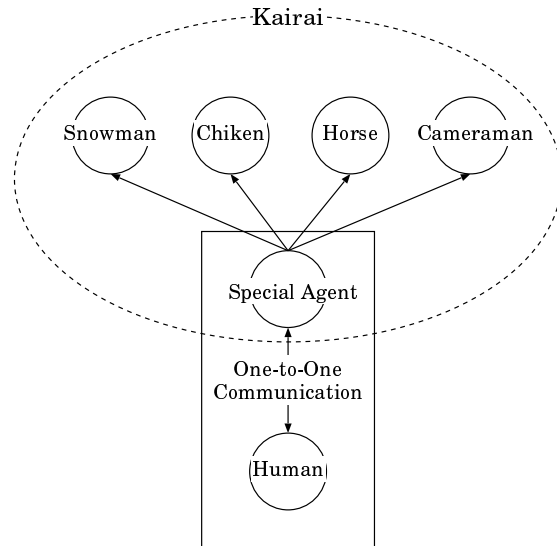


Figure 3: One-to-One Communication in Kairai

Since current software robots in Kairai are not autonomous, it is very difficult to conduct cooperative actions including several robots. Consider “gazing” (Rickel et al., 2001), a simple cooperative action. In the current Kairai system, even though a software robot is conducting a task in the virtual space, the other

robots are not paying any attention to these actions. In human society, it is natural for a person to gaze at another one working near him/her. These were already implemented in (Cassel et al., 1999), but due to the absence of autonomous robots in current version of Kairai, it is impossible for any robot to directly communicate with another. Problems of gazing as well as the other cooperative actions can be solved by introducing autonomous robots and one-to-many conversation mode in a virtual space. The latter will be discussed more in the next section.

Currently, Kairai does not deal with non-verbal phenomena including intonation in speech, gazing, facial expressions, body actions including hand gestures. Facial expressions are related to emotional actions. As para-linguistic phenomena play important role in communication, we would like to account for para-linguistic phenomena as one of challenging research topics in the future. Fortunately, compared to hardware robots, software robots emulate para-linguistic behavior much easier since they do not have any mechanical limitations.

6 One-to-Many Conversation

In addition to the items mentioned in the section 2, NLU systems should deal with a one-to-many conversation in addition to one-to-one conversation. One-to-many conversation makes sense in the multi-agent environment (Febler, 1999) (Weiss, 1999), since in the one-to-one conversation it is easy to decide who the intended listener is. On the contrary, in the one-to-many conversation as shown in 4, as there are many potential listeners, it is difficult to decide to whom a command issued by a speaker is intended. Usually, the listener is mentioned explicitly in the first dialogue, but that he/she is not mentioned in the rest of the dialogue. Confusion can happen among agents if each agent is unable to recognize who is the actual intended agent who needs to perform some tasks according to the command given by a speaker. The problem takes place when a subject or an object does not appear in a command due to ellipses.

To understand the above problems clearly, consider the following conversation in a multi-agent environment.

1. Human: "Hey, Robot A, I am throwing a red ball." >Robot A looks at Human.

2. Human: "Catch it."

>Robot A begins the action to catch the red ball.

Note that the above last Human command lacks a subject, but Robot A has to perform the action "catch" along with resolving the anaphora expression "it" that should correctly be identified as the red ball. The other robots should not perform the action "catch" even though they can hear "Catch it" command.

It seems obvious that each software robot should be autonomous in the multi-agent environment and have the ability to control his behavior in his own right.

7 Conclusions

After reviewing an NLU system in the past, we pointed out several important issues concerning the building of the next generation NLU system. Kairai, a system which was developed at Tokyo Institute of Technology, played a key role in exemplifying the issues mentioned in the preceding sections.

Even though the system under consideration is composed of a set of software robots, the research results are applicable to any future multi-agent system consisting of hardware robots. We have also discussed the importance of para-linguistic phenomena in addition to emphasizing the need for better algorithms for anaphora resolution and ellipsis handling. Additionally, dealing with ill-formed sentences which frequently occur in spoken language is also an important issue requiring attention.

The next generation NLU system ought to be a multi-agent system, with a wide array of application areas such as:

1. Entertainments
2. Helper robots (medical and in-home use)
3. Tutoring systems
4. Sign language systems,
5. Navigation in a virtual space,
6. Electrical Appliances,

Readers are easy to understand the application areas of entertainments and Helper robots, although the latter will needs higher and more

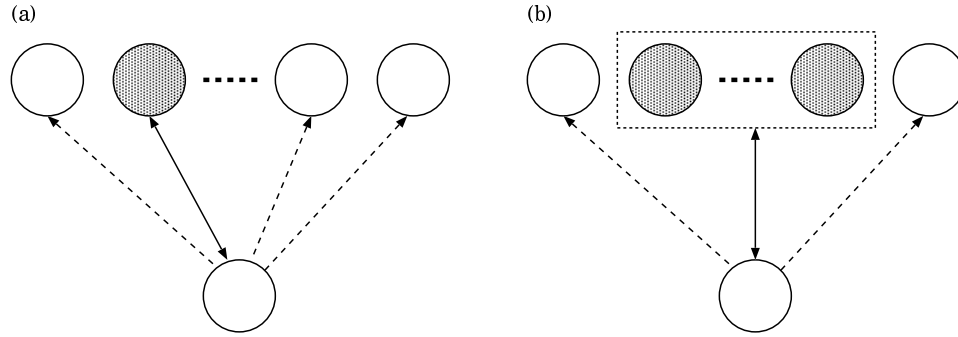


Figure 4: One-to-Many Conversation

reliable technologies in the future. Instead of a written manual we have to read, it is better for a software robot (or a pedagogical agent) to instruct us a method how to manipulate devices by showing its 3-D models.

The navigation with voice commands in natural language is possible to apply to any kind of virtual space. Typical examples are 3-D models of internal organs, molecular structure, DNA structure, and geographic space etc.

Let us expand on the final item. The future electrical appliances will be equipped with “ears” for listening to user’s commands and will need to process these commands to execute them similar to current software robots. In such circumstances, the research on both multi-agent system and one-to-many conversation system will become increasingly important.

Acknowledgments

Thanks to Dr. Taiichi Hashimoto and Mr. Slaven Bilac for their help to prepare this manuscript.

References

- N. I. Badler, C. B. Phillips, and B. L. Webber. 1993. *Simulating Humans - Computer Graphics Animation and Control*. Oxford University Press.
- N. I. Badler, M. S. Palmer, and R. Bindinganavale. 1999. Animation Control for Real-Time Visual Humans. *Comm. of the ACM*, pages 65–73.
- N. et al. Badler. 1998. A Parameterized Action Representation for Virtual Human Agents. In *Proceedings of the 1st Workshop on Embodied Conversational Characters*, pages 1–8.
- J. Cassel, T. Bickmore, L. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea-

In *Proceedings of CHI’99 Conference*, pages 520–527.

- P. R. Cohen, J. Morgan, and M. E. Pollack, editors. 1990. *Intention in Communication*. The MIT Press.
- J. Febler. 1999. *Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence*. Addison-Wsley Longman.
- B. J. Grosz. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.
- J. Rickel, Ruth Aylett, and Daniel Ballin, editors. 2001. *Intelligent Virtual Agents for Education and Training: Opportunities and Challenges*. Springer.
- Y. Shinyama, T. Tokunaga, and H. Tanaka. 2000. “Kairai” - Software Robots Understanding Natural Language. In *Proceedings of the 3rd Workshop on Human Computer Conversation*, pages 158–163.
- Y. Shinyama, T. Tokunaga, and H. Tanaka. 2001. A Software Robot Kairai that Understands Natural Language (in Japanese). *Journal of JIPS*, 42(6):1359–1367.
- H. Tanaka. 2000. *Language Understanding and Action Control*. Ministry of Education and Science.
- H. Tanaka. 2001. *Language Understanding and Action Control*. Ministry of Education and Science.
- G. Weiss, editor. 1999. *Multiagent Systems*. The MIT Press.
- T. Winograd, editor. 1972. *Understanding Natural Language*. Academic Press.