ICOT Technical Memorandom: TM-1249

TM-1249

対話管理システム ToR '92

丸山　友朗、鈴木　浩之、
飯塚　泰樹（松下）

March, 1993

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5

Institute for New Generation Computer Technology

Discourse Management System ToR'92

Tokyo Information and Communications Laboratory
Matsushita Electric Industrial Co., Ltd.

1992-12-07

1 Introduction

In the future computers will understand human speech, and we would like to contribute towards this goal by establishing natural language processing technology. This technology must have the following characteristics:

1. **situatedness,**
2. **robustness**, and
3. **incrementality.**

Utterances should be interpreted differently depending upon the situation and manner in which they are spoken. This is called the "situatedness" of language. To give computers this ability, our research is aiming at how best to represent these situations. We have developed a new knowledge representation language based on Situation Theory, and incorporated it into ToR '92.

One reason natural language interfaces are better than menu based interfaces is that the user can input anything. This means that the system must somehow handle all kinds of inputted sentences. So we can identify many requirements to dialog system. Amongst them, we concentrate on

1. how to segment sentences into words, and
2. how to identify and represent the unknown words.

One big defect of natural language interface is that it is too slow. We consider a major reason for this to be that machines start processing after an entire sentence has inputted, while humans begin processing utterances as soon as the first word is heard. Our system tries to mimic this human ability by taking an incremental approach to processing word sequences.

ToR'92 consists of four parts:

1. Utterance analysis part,
2. Discourse management part,
3. Utterance generation part, and
4. Knowledge representation part.

When a sentence is input into ToR'92, the utterance analysis part analyses it along with visual and phonic information. The utterance analysis part generates representations of the interpretation of the utterance as an output.

The discourse management part first stores the pair of representations, namely the representation of the interpretation and the representation of the situation, to keep track of the utterance sequence. Then it tries to prove whether the interpretations are valid in this world. The proof procedure and its result are passed to the utterance generation part.

The utterance generation part generates a sequence of appropriate words, the content of which comes from the discourse management part.

The knowledge representation part is the basis of ToR'92. This part provides basic utilities that are needed to construct the other three parts, and includes utilities for uniform treatment of information from various sources.

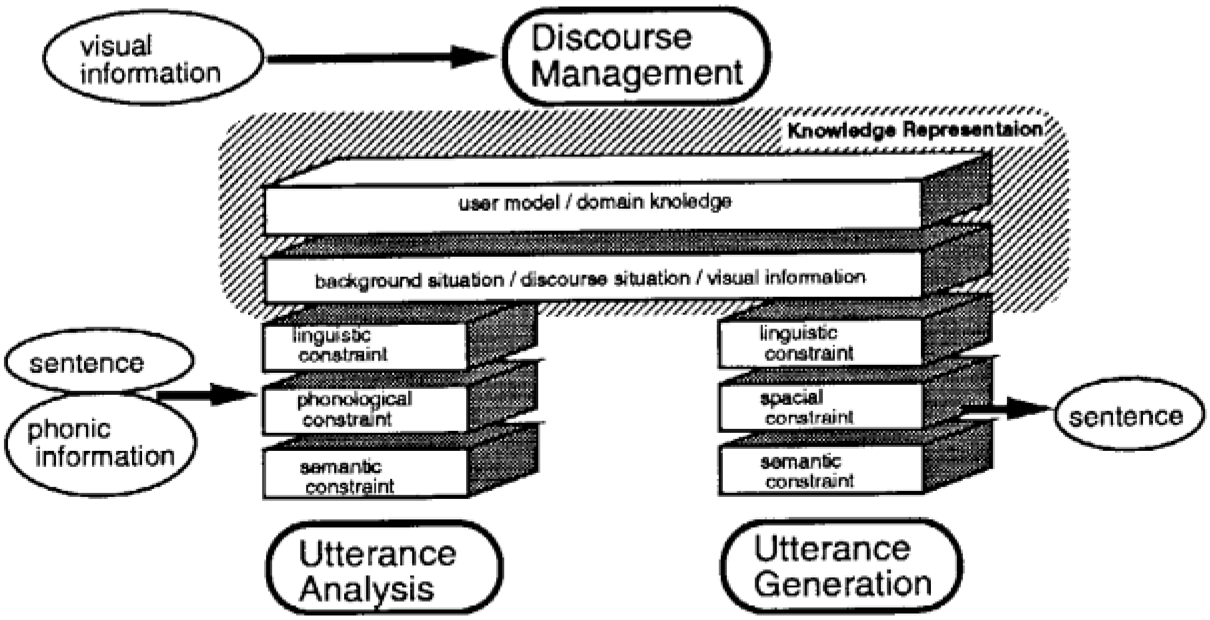

Figure 1: The system architecture of ToR'92

2 Utterance Analysis

The utterance analysis part produces the meaning representations of Japanese sentences inputed by a user. It consists of three kind of analyzers, the morphological analyzer, the syntactic analyzer, and the semantic analyzer.

2.1 Morphological Segmentation

As you may know, Japanese sentences are written without separating words with spaces. For example, "Tanaka san wa imasuka" will be written "田中さんはいますか". However, we need to identify what words are used in order to process the sentence. "田中さんはいますか" should be somehow separated into a word sequence like "田中 さん は い ます か".

The morphological analysis subpart does this task as the first step of analyzing the utterance. Dividing sentences into the appropriate words,however,is not so easy to do. See the following.

"かのじょがくるまでまちます"

↓

"かのじょ が くる まで まち ます"
彼女 が 来る まで 待ち ます

(I will wait until she comes)
OR
“かのじょ　が　くるま　で　まち　ます”
彼女　が　車　で　待ち　ます
(She waits in a car)

This process, as you see, utilizes the dictionary heavily. To accept more sentences, a larger dictionary is needed. However if the dictionary becomes larger, the search requires more time, and hence the system becomes slow time and is not comfortable as a human interface.

We are trying to speed up the dictionary consulting process, by improved the data structure kept in the computer at the processing stage. Wa are doing this by improving the way the dictionary data is kept in computer.

2.2 incremental analysis

Human beings grasp the partial meaning of an utterance while hearing the utterance word by word. ToR'92 tries to mimic this human ability by incorporating an incremental parsing method. That is, syntactic analyzer constructs a parse tree incrementally upon a word is inputted. However it is hard to decide whether a given word sequence will make a phrase without the help of further information. We utilize phonic information to determine this timing. When syntactic analyzer gets a pause symbol, it resolves existing syntactic ambiguity and grows up the parse tree. This method is supported by rules describing the relationship between syntax and intonation elucidated by text speech synthesis.

2.3 Analysis using heterogeneous information

Generally speaking, sentences are very ambiguous while utterances are not. In our everyday life, it is very rare when we notice an alternative interpretation of a sentence uttered, while the syntactic analysis of a sentence shows that in fact there are five to ten possible interpretations of it.

This is (partly) because we know each other, share lots of information other than the sentence used, and can utilize this background to interpret the utterance. Among this contextual information, visual information and phonic information are closely connected to utterance.

- Using phonic information

  Given data of intonation for the input sentence through symbols inserted in the sentence which correspond to pause, accent, and pitch, the system uses them to reduce various ambiguities. We show some examples in the following.

  - When the pitch of the sentence end is up, the syntactic analyzer can infer that the sentence may be interrogative.
  - We pronounce one meaning clause in one breath. In syntax, words pronounced by one breath are strongly connected. So, using this rule,the syntactic analyzer may select the correct one in many situations of syntax ambiguity.
  - Japanese words have at most one accent. Using this rule, the morphological analyzer can reduce the ambiguity involved in breaking a sentence down into words.

- Using visual information

  Visual information is used to process demonstratives.

  For example, Japanese “これ”(“this”) is used to point an object near the speaker, and Japanese “あれ”(“that”) is used to point an object far from the speaker. Using this rule and visual information, we can do the above process.

3 Situation management

The situation management part manages almost all the information which is held in this dialog system, and provides some methods to access the informations for other parts. This part also plans the system's action to determine the system's response.

This part consists of two modules, the situation management component and the planning component.

Situation management module classifies informations into the appropriate situation and manage this information. Planning module plans the system's action to verify propositional contents of the user's utterance, and hands the result of the action to the natural language generation part.

3.1 The method of situation management

To manage a lot of information, our system has some situations, and classifies information which comes from other parts as the content of particular situation. Each situation has its own role, such as tracking the meaning content of the utterance. If system need to refer particular information, it searches the appropriate situation and get the information.

The system has situations as follows (Fig 2) :

- *Contextual information*

  This situation holds informations about when, who, to whom, and utterances. This information is referred to when the context information is needed.

- *Circumstantial model*

  This holds information about the circumstance in which the utterance occurred. It includes visual information.

- *System's knowledge*

  This holds information which comes newly into the system. It includes the result of the system's action.

- *User model*

  This holds informations about user. It includes the user's social standing.

These situations are subsituations of situation called "the system model". And the situation management part provides several methods to directly access to this information for other parts.

3.2 Useing heterogeneous informations

In ordinary dialog, we use not only text information such as the contents of utterances, but also many other information including visual information. One of our system's aims is to handle this heterogeneous information at same level as text information. In paticular, we give priority to visual information.

In our system, visual information is showing as picture on the display (Fig 3). Information we can get directly from this picture is geometric information (x-y axis). But if we need to use visual information, we have to convert the geometric information to semantic information.

For example, from Fig 3 we can get information such as X's position (250,200) and Y's position (450,230). The system converts this information to:

$$\langle\langle left\_of, (object1 : X, object2 : Y), 1\rangle\rangle$$

Using expressions like these, the system can understand which object is referred to by the expression "that" in the sentence "Is that man Mr.Tanaka?".

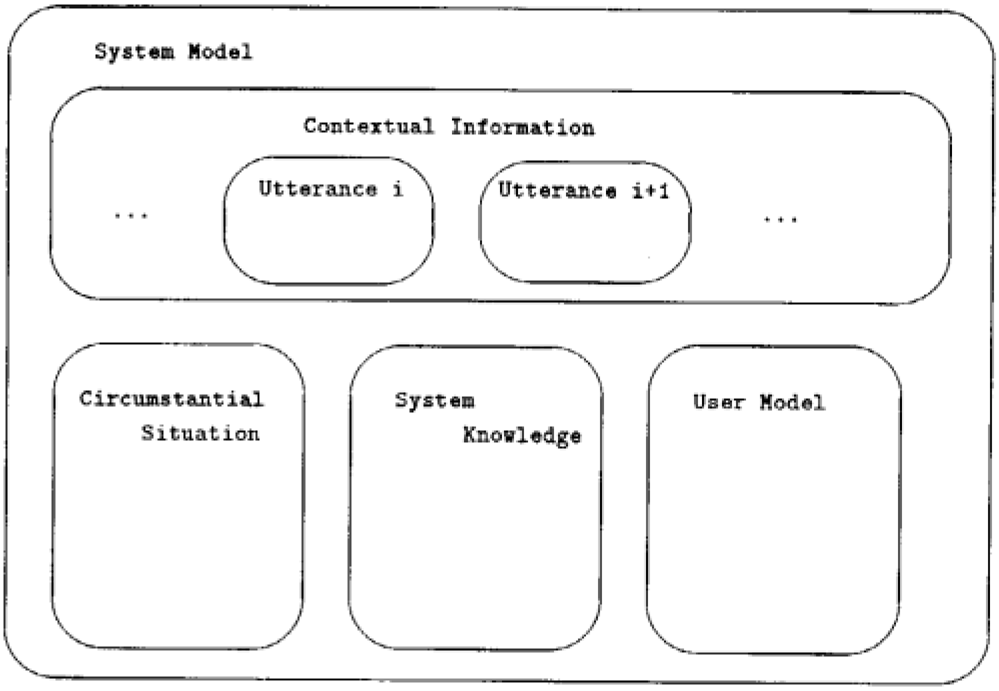

Figure 2: Construction of situations in our system

3.3 Planning module

The planning module plans the system is verication of the content of the utterance. For example, if the propositional content of the utterance is as the following formula:

$$\langle\langle present, (agent: \text{"}tanaka\text{"}, place: \text{"}here\text{"}, time: \text{"}tomorrow\text{"}), 1\rangle\rangle$$

The system starts to verify whether Mr.Tanaka will be here tomorrow. To verify this, system will search his schedule and check it, or ask Mr.Tanaka about his schedule for tomorrow if he is present. The planning module determins which action the system should do to verify the content of utterance.

4 Utterance generation

4.1 Introduction

Natural language generation is the deliberate production of a sentence or text to meet some communicative goals of a speaker. It consists of the following major activities:

1. determining what information is to be uttered
2. imposing a suitable order on the elements of the information consistent with the constituent structure of language and expressing the relative salience and newness of the elements.
3. determining what wording and syntactic constructions to use.

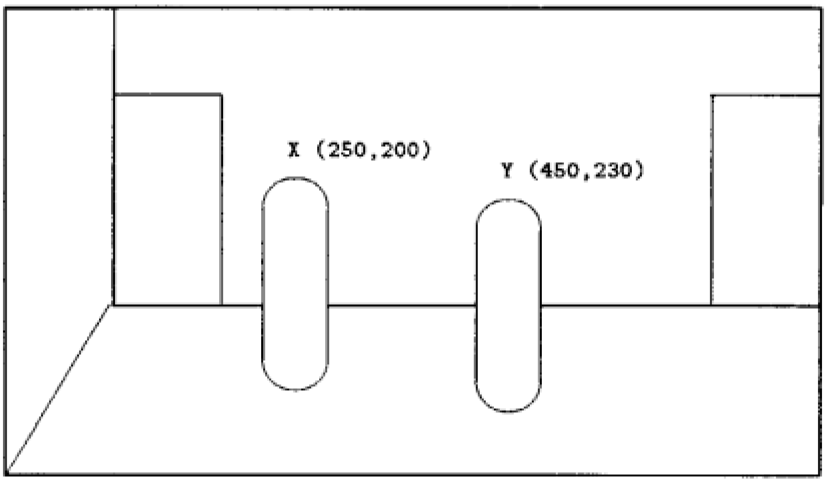

Figure 3: Visual information on the display

These activities are divided into two generation processes. The first two activities are called "text planning" or deciding "what-to-say ". The third is called "realization" or deciding "how-to-say".

We have been studying both od these processes since the ICOT project started.

As for problems with the realization process, we use some result form Japanese linguistics research and logic programming research to solve them:

- grammatical notation which is suitable for Japanese
  unification grammatical formalism

- knowledge representation for describing rules
  CIL programming language (PST — partially specified term)

- efficient mapping mechanism from a knowledge representation to a grammatical representation
  ESP(Prolog like language) and CIL

As one of the results of our research, we developed a software package named "the Language Tool Box(LTB)", which provides a function for generating a sentence from a given intermediate representation.

By using the LTB software package for generation, one can make (Japanese) natural language processing software easily and quickly. Part of the software modules of the discourse management system ToR is constructed on the LTB generator.

Next,regarding to research on text planning, it is a more complex and ill-formed process compared to the realization process.

To plan text with the competence of a human being, it is not sufficient to have only descriptions of the syntactic,semantic, and discourse rules of a language: human language behavior is part of a coherent plan of action directed towards satisfying a speaker's goal. Furthermore, what a speaker will utter is dependent upon not only the linguistic constraints but the situation in which the speaker is embedded.

The difficulties of producing the entire range of utterances that human beings can generate led us to restrict types and domains of discourses with which the system deals.

At present, the type of discourse which the system can produce is a task-oriented dialogue, for which we have developed natural language generation system which satisfied some requirement. We can identify a number of requirements of the system. They are as follows.

- **cooperativeness**
  The system should have the ability to produce cooperative utterances in a task-oriented dialogue.

- **real time**
  The system should perform its tasks in real time.

- **incrementality**
  The system should be able to generate not only a whole sentence but also sentence fragments.

- **robustness**
  The system should be robust.

To satisfy these requirements, we have been developing new frameworks for planning sentences, which were applied to the experimental systems, ToR'90 and ToR'92.

In the next chapter, we briefly explain the new generation strategy adopted in the ToR'92 system.

4.2 Generation framework of the ToR'92 system

In the framework on which ToR is based ,the production of sentences by the system is considered a kind of action like; as ordinary physical actions, the production of sentences is an entailment of performing some plan in mind.

Thus, to control the process of producing utterance and deciding what sentences should be uttered, we can use a model of th emid in which are settled some proper axioms.

The axioms used in the ToR'92 system are the following:

Axiom1 An utterance is a report of the mental state of an agent.

Axiom2 When an agent hears an utterance, he tries to verify its (propositional) contents.

Briefly, Axiom 1 means that a human utters what comes into his mind. A cognitive action of an agent has various effects on his mental state. When some changes are caused in his mental state he can produce sentences or sentence fragments as reports of his mind.

Axiom 2 is needed to connect an utterance of the system to an utterance which another agent produced just before.

Although the axioms looks like very simple, they are sufficient to produce the following basic types of discourse elements:

- facility of direct speech act
  - Question-Answer pair
    (example)

```
Q.   田中さんは、   明日、    いる？
     Mr.tanaka      tommorow  will-be ?

Ans. はい、田中は、明日、    ここに、います。
     Yes  Tanaka  tomorrow  here    will-be
```

- Asking-Execution pair

```
Q.   東京行きの切符を、  買ってください。
     Tickets for Tokyo   please buy

Ans. はい、買いました。
     Yes   bought
```

- facility of indirect speech act

```
Q.   田中さんに、    電話できるかな
     To Mr.Tanaka    can I phone

Ans. 田中は、ICOTに出張ですね
     Tanaka  make a busines trip to  ICOT
     電話番号は、      123-4567 です。
     The phone number  123-4567 is
```

For example, the system can produce an "answer" to a "question" of another agent along the following steps:

step1 "Question" is a sentence which contains a propositional content," The speaker doesn't know the truth value of a proposition P". The system get the meaning representation of proposition P by analyzing the sentence.

step2 the system tries to verify P according Axiom 2. To verify it, the system constructs a verifying plan.

step3 When the system performs the plan, the result reflects the mental state of the system. According to Axiom 1, any changes caused in the mental state are translated into sentences as reports of it.

The Axioms are so simple and applicable to any input sentence that the system becomes robust and perform given tasks very quickly. Furthermore, the sentence which are produced are helpful for participants in the dialogue.

Also we adopted other axioms and "constraint filters" to make the response of the system more sophisticate. The ToR'92 system has 2 kinds of constraint filters, which can eliminate unimportant utterances from complete utterances that are produced.

- **temporal filter**
  interruption of production of a sentence which takes a shorter time to produce than a given threshold time

- **syntactical filter**
  eliminate some grammatical constituents(Bunsetu) that have simular syntactical shapes compared with constituents in a previous sentence.

It is characteristic of the ToR'92 system that the filters are independent from any semantical information,which are often used in other dialogue systems and make their temporal efficiency very low.

5 Knowledge representation language

Almost all information that is held in this system is described in a particular format called"Knowledge representation language".

Knowledge representation language provides the format in which much information is described and methods to clasify this information. In our dialog system, we use "$\mathcal{LAST}$" – semantic representation language based on the idea of "Situation Semantics".

For example, the meaning of the sentence "Mr.Tanaka will be here tomorrow" is represented as following formula.

$$\langle\langle present, (agent: \text{"}tanaka\text{"}, place: \text{"}here\text{"}, time: \text{"}tomorrow\text{"}), 1\rangle\rangle$$

This formula represents the relation that"*present*" holds between three arguments (*tanaka, here, tomorrow*). And this information is kept in an appropriate "situation".

$\mathcal{LAST}$ has many functions, such as creating these representations, searching the information among many situations, and so on.

6 research themes in the ToR'92 system

In the ToR'92 system, we try to provide the following major facilities:

- incrementality
  Sentence fragments(ie. incomplete sentences) can be produced.

- attitude
  The ToR'92 can't handle Japanese attitude auxiliary verbs.