Scene Direction Based Reference In Drama Scenes

Hiroshi Nakagawa
Yokohama National University
79-5 Tokiwadai, Hodogaya
Yokohama, 240, Japan
nakagawa@naklab.dnj.ynu.ac.jp

Yoshitomo Yaginuma **Masao Sakauchi**
Institute of Industrial Science
University of Tokyo
Roppongi, Minato, Tokyo, 106, Japan
{ yaginuma, sakauchi }
@sak.iis.u-tokyo.ac.jp

Abstract

Our research target is reference relations between descriptions of script and an actor/actress who actually plays in the drama scene that correspond to the scene direction which is a part of the script. In this paper, first we analyze sentences used as the scene directions, and classify them. Then we propose the rules to extract subjects and predicates from those sentences. With the extracted subjects and predicates, we build the existence/action map that explains the situations happening on each scene. The existence/action map we build describes scenes very correctly as for whether each player appears in each scene or not. Our experiment shows that the recall is around 80% and the precision is over 90%. This means that our system of inferring reference relations works well for scene directions. Then we develop the scene retrieval system in which this map is used to retrieve scenes from the drama video database according to the input query. We also show some experimental results of our retrieval system.

1 Introduction

In conventional multimedia retrieval systems, image data are indexed manually because image processing technologies have not yet provided us with image understanding methods that are powerful enough to automatically extract useful indices from image data themselves. Then, in order to enhance multimedia retrieval systems, we have to employ other types of information source. That is why we focus here on drama scripts that are apparently natural language media. However it is obvious that only using scripts does not guarantee effective retrieval. The essential point is, of course, the combination of information from language resource and image data. This combination can drastically improve the quality and efficiency of multimedia retrieval. From this point of view, reference relations between a description of script and an actor/actress who actually plays in the drama scenes that correspond to the script are the most useful pieces of information for scene retrieval, and consequently they become our research targets. If we successfully identify this kind of reference, it is a great help for multimedia information retrieval, especially scene retrieval from data base of drama videos. However the whole script of drama is so complicated that we cannot deal with by today's natural language processing technologies. Then we concentrate our focus here on **scene directions** that explain the situation of each scene in the drama, for instance, actors/actresses' position, movement, and so on.

In section 2 we describe the pieces of information which describe a scene. In section 3 we explain our natural language processing system that extracts information describing each scene. There we also show the experimental results of our natural language processing system. In section 4, we show some results of scene retrieval system. Section 5 is the conclusion.

2 Scene Descriptions

Each scene of a drama is characterized by the following five types of information.

1. Location of the scene.
2. Time of the scene.
3. Players on the scene. Actually role name of each player is described in the scene direction.
4. Physical or psychological states of each player in the scene.
5. Actions of each player in the scene.

These types of information are described in the scene direction for individual scene. States and actions of a player are described basically for each utterance or action in the scene direction. Henceforth we call a time unit corresponding to an utterance or an action as *sub-scene*. As we will show later, a subscene is a unit of retrieval. Then we use the following classification in order to describe each player's status

on each sub-scene: 1) being absent (ABS), 2) existing(EXI), 3) conversing(CON), and 4)acting(ACT). *Acting* is further described by the verb used in the scene direction sentence. Using this classification, we can express the contents of scene such as shown in the table 1 which we call *existence/action map* of scene. The left most column expresses a sequence number of sub-scene in the scene. The second, third, fourth and fifth columns correspond to the status of each player whose name is shown in the fourth raw. The sixth column describes the detail of action by the corresponding verb.

Table 1: An example of existence/action map

scene 1					
Location: Police office					
Time: evening					
sub scene	Alice	Betty	John	Bill	action
0	ABS	ABS	EXI	EXI	
1	ACT	ACT	EXI	EXI	coming back
2	EXI	CON	EXI	EXI	conversing
3	EXI	EXI	CON	EXI	conversing
4	EXI	EXI	ACT	EXI	going out
5	EXI	EXI	ABS	EXI	

This example of existence/action map is interpreted as follows. At sub-scene, 0 John and Bill exist there. At sub-scene 1, Alice and Betty come into this scene. Then at sub-scene 2 and 3, all four persons are there, and Betty and John speak one after another. At sub-scene 4, John goes out from the scene. Therefore at sub-scene 5 he is no more in the scene. This kind of map is used in retrieving the image data of sub-scene as later described.

3 Scene Directions Analysis

3.1 Sentence Patterns of Scene Directions

In this section, we describe how to extract information from scene directions in order to build an existence/action map of the scene. For this purpose, we first characterize the scene directions that are actually restricted Japanese sentences. Simple sentences used as a scene direction are classified into the following six patterns. We also show an example sentences of each pattern:

1. **subject , verb phrase**

(1) Taroo to Hanako ga
and NOM
kaette -kuru.
come back -kuru
'Taroo and Hanako come back.'

In this type of sentence, "ga"(subject marker), "wa"(topic marker) or "mo"(topic marker + 'too') are used as a nominative particle. Moreover "ga" and "wa" are sometimes replaced with a comma ",".

2. **verb phrase , subject**

(2) Soto o miru Taroo.
outside ACC see
'Taroo who sees outside.'

3. **verb phrase**

(3) Odorite iru.
surprised being
'ϕ is surprised.'

4. **subject, noun phrase**

(4) Taroo ga hitori.
NOM alone
'Only Taroo is there.'

5. **noun phrase, copula**

(5) Denwa dearu.
phone call COPULA
'A phone call arrives.'

6. **noun phrase**

(6) Sibaraku-no tinmoku.
for a while silence
'No one speaks for a while.'

We employ a very simple pattern matching based information extraction system described later based on two reasons: 1) The structures of simple sentence used as a scene direction are limited within these six patterns. 2) What we would like to extract from scene directions in order to build an existence/action map is only the following two references. Namely who is the subject, and what action or state the referent of the subject does or is in. Therefore it is enough to extract the subject and the verb (or 'the verb + the auxiliary verb').

3.2 Subject Extraction

Subjects are extracted by matching the patterns S generated by the following rules.

rule 1 P is a proper name or a common noun.

rule 2 $P \leftarrow P$ ","P

rule 3 $P \leftarrow P$ "to "P, where "to" means 'and' in English.

rule 4 $S \leftarrow P$ ","| P "ga"| P "wa"| P "mo"
Here a subject corresponds to P. This rule is for pattern 1 and 4.

rule 5 $S \leftarrow P$ "."
This rule is for pattern 2.

rule 6 $S \leftarrow P$ "dearu" | P "da", where "dearu" and "da" are copulas in Japanese.
This rule is for pattern 5.

It is not necessary to extract a subject from the sentence of pattern 6. A sentence of this pattern usually describes the atmosphere of the scene. As for pattern 3, we have to infer the referent of omitted subject, namely *zero subject.* We have a plenty of theories for this purpose including centering theories (Brennan et. al , 1987; Kameyama, 1988; Walker et. al , 1994). Here, however, we employ a very simple rule as follows.

rule 7 The referent of zero subject is the same as the referent of subject of the previous sentence.

This rule is a small subset of centering theory, but as you will see later, it works well to extract a subject from a sentence of scene direction. We also apply this rule for a complex sentence in which a subject of main clause is omitted. Namely the omitted subject is deemed to corefer with the explicit subject of subordinate clause. The reason of this is that 1) in a scene direction, a sequence of actions is described, and 2) in a complex sentence of scene direction, a subordinate clause describes an action or state that happens prior to the action or state described by the main clause. In other words, a subordinate and a main clause of the complex sentence can be regarded as two consecutive simple sentences.

3.3 Predicate Extraction

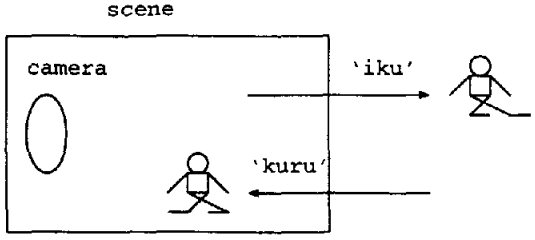
As for pattern 1, 2 and 3, we can extract a predicate just by extracting a verb or a verb + an auxiliary verb from the sentence. In pattern 4, we cannot identify the exact action or state from the sentence. But at least, we know that there exists a person the subject refers to. Therefore a predicate extracted from a sentence of pattern 4 is regarded as "exist" by default. We couldn't find any reasonable predicate for sentences of pattern 5 and 6. Then we also use "exist" in these patterns as we do in pattern 4. In sum, we use the following rule to extract a predicate from a scene direction.

rule 8 A predicate of the sentence is either a verb (+ an auxiliary verb) used in a sentence of pattern 1, 2, or 3, or "exist"in a sentence of pattern 4, 5 or 6.

3.4 Building An Existence/Action Map

In this system, the purpose of semantic interpretation is limited to build an existence/action map from the extracted subjects and predicates by the way described in section 3.2 and 3.3. For this purpose, the key element of predicate is the so called directional auxiliary verb(Kuno, 1978). There are several directional auxiliary verbs in Japanese. Among them, the most essential ones for our purpose are "-tekuru" and "-teiku". The directions indicated by these auxiliary verbs are defined relative to the position of camera. If "-tekuru" is a part of the predicate of sentence, a referent of the subject of the sentence comes into the camera angle and/or is approaching the camera. If "-teiku" is a part of the predicate of sentence, a referent of the subject of the sentence goes away from the camera and probably is out of the camera angle. The situation is depicted as shown in figure 1.

Figure 1: Movement of "tekuru" and "teiku"



In addition, basic verbs, "kuru('come')" and "iku('go')", also express the same type of sense of direction as "-tekuru" and "-teiku" do respectively. By these considerations, we drive the following two default rules to infer existence or absence of the referent of the subject in the scene.

rule 9 If the predicate of sentence includes an auxiliary verb "-tekuru" or a verb "kuru", a referent of the subject of the sentence had not been in the scene beforehand, and just has come into the scene.

rule 10 If the predicate of sentence includes an auxiliary verb "-teiku" or a verb "iku", a referent of the subject of the sentence will not exist in the scene afterward.

Of course, these two rules are default rules, and there are exceptional cases. For instance, if a sentence explicitly describes that Taroo has existed in the scene beforehand, even though we encounter the sentence that

(7) Taroo , tikayot -tekuru.
SUB approach
'Taroo approaches to here.'

, we infer that Taroo has already been in the scene. Or in "-teiku" case, if the sentence:

(8) Taroo, Hanako -ni tikayot -teiku.
SUB -DAT approach
'Taroo approaches to Hanako.'

is followed by the sentence:

(9) Taroo , Hanako -ni
SUB -DAT
nagur -areru.
hit -PASSIVE
'Taroo is hit by Hanako.'

, then we infer that Taroo is still in the scene after the action described by (9).

Another exceptional case is that a verb is either stative or state change without action. For instance, the sentence:

(10)
Taroo , kaoiro-ga-warukunat -teiku.
SUB become pale
'Taroo turns pale.'

indicates that Taroo is still in the scene when he looks pale. We can identify these kind of verbs, stative and state change without action, with a dictionary like IPAL-Basic Verb Dictionary For Computer(IPA, 1990).

Although the rules and the treatment of exceptional cases are very important, generally if a sentence of scene direction describes that the referent of subject does an action or is in a certain state, the referent surely exists in the scene. Another group of expressions that are frequently used and are important to build an existence/action map are a stative verb "iru('exist')" and its negation "inai('not exist')." They explicitly show the existence or non-existence of a referent of subject in the scene. Then we have the following two rules.

rule 11 If a sentence describes an action or state, the referent of the subject is in the scene.

rule 12 If "iru" is used as a predicate, the referent of the subject is in the scene.
If "inai" is used as a predicate, the referent of the subject is not in the scene.

These two rules can override the results we infer by rule 9 or rule 10, because rule 9 and 10 are default rules and rule 11 and 12 explicitly describe the scene. One question we have here is that if we encounter a negative predicate like "inai", how should we infer. However, in reality, we don't find negative predicates except for "inai" in scene directions, because scene directions describe what players should do in the scene, and they almost never describe what players shouldn't do in the scene. Things not to be done in a scene are usually directed by the human director of the drama.

Now we show an example of existence/action map built from the following scene directions.

(11) Taroo to Jiro ga
and SUB
kaet -tekuru.
come back
'Taroo and Jiro come back.'

(12) Taroo , " tadaima "
SUB " I'm home "
'Taroo says " I'm home."'

(13) Hanako , " gokurousan "
SUB " you did well. "
'Hanako says " You did well"'

(14) to nagusameru.
and comfort
'and comforts two of them.'

(15) Saburoo , tokei -o miru.
SUB clock -ACC look at
'Saburoo looks at the clock.'

(16) Saburoo , de -teiku.
SUB go out
'Saburoo goes out.'

The existence/action map derived from (11) through (16) is as follows.

sub-scene	Taroo	Jiro	Hanako	Saburoo
	action			
before 11	ABS	ABS	EXI	EXI
11	ACT	ABS	EXI	EXI
	come back			
12	CON	EXI	EXI	EXI
	speak			
13	EXI	EXI	CON	EXI
	speak			
14	EXI	EXI	ACT	EXI
	comfort			
15	EXI	EXI	EXI	ACT
	look at a clock			
16	EXI	EXI	EXI	ACT
	go out			
after 16	EXI	EXI	EXI	ABS

We build an existence/action map by the following procedure.

step 0 Step 1 through step 3 are applied sequentially for scene directions in a sentence by sentence manner.

step 1 A sentence of scene directions is analyzed with the Japanese morphological analyzer JUMAN(Matsumoto et. al , 1992) to segment a sentence into a sequence of word accompanied by part of speech tags.

step 2 The subject and predicate of sentence are extracted using rule 1 through 8.

step 3 For each player the value of sub-scene, namely, ABS, EXI, CON or ACT, is inferred with rule 9 through 12.

We did build existence/action maps for scene directions of five Japanese dramas. These include a suspense drama, a home drama, a love story, a school life drama, and a comedy drama. Each drama lasts one hour (including CM time). The number of the sentence we analyze is 1272. The first results which are shown in the table 2 are the rates that step 1

and 2 correctly extract subjects and predicates. We use not a parser which is based on phrase structure rules but a simple pattern matching based on rule 1 through 8. Nevertheless these results indicate that our rules for extracting subjects and predicates work quite well.

Table 2: Rate of correctly extracted subjects and predicates

	correct subjects	correct predicates
drama 1	94.9%	98.7%
drama 2	82.4%	93.1%
drama 3	85.8%	94.8%
drama 4	75.7%	94.9%
drama 5	83.1%	97.2%

The main reason of failing to extract a subject is the failure of inferring the referent of zero subject. The almost all reasons of failing predicate extraction is the failure of morphological analyzer.

The next results we show are the accuracy of our existence/action map. The key factor for scene retrieval is whether a specific player appears on the scene or not. Therefore we focus on how accurately existences and absences, namely EXIs and ABSs, are inferred. We estimate this with recall and precision rates defined as follows.

$$recall = \frac{\#\mathrm{CI}}{\#\mathrm{C}}$$

$$precision = \frac{\#\mathrm{CI}}{\#\mathrm{I}}$$

where $\#CI$, $\#I$, and $\#I$ means "number of correctly inferred cases in the map by our rules", "number of cases to be correctly inferred" , and "number of all cases inferred by our rules" , respectively. The results are shown in table 3, and they are extremely encouraging ones.

Table 3: Recall and precision of existence/action maps

	Recall	Precision
drama 1	86.7%	97.8%
drama 2	73.2%	96.8%
drama 3	91.7%	91.7%
drama 4	87.9%	93.5%
drama 5	77.2%	94.4%

Our rules derived based on semantics of "-tekuru" and "-teiku" are proven to work correctly in almost all cases. The remarkable point in these results is that even though our natural language analysis system employs a shallow understanding mechanism which is easily implemented with today's NLP technologies, it works very well for scene directions. This is a very limited area but useful for scene retrieval system, which is a promising application of multimedia information retrieval.

4 Scene Retrieval System

We develop the scene retrieval system for drama scenes based on an existence/action map. A sub-scene is a unit of retrieval because players on the scene change in a sub-scene by sub-scene manner as you see in an existence/action map. Therefore we have to find the correspondence between a sub-scene and a set of real image frames. The multimedia data we have consist of 1) a sequence of image frames, 2) audio track data, and 3) script of drama including scene directions. The temporal correspondence, or in other words *synchronization*, of these three types of media data is calculated with DP matching technique we have already developed(Yaginuma, 1993). Owing to these correspondences, we identify each set of frames that corresponds to the each part of dialog. Then we regard a set of sequential frames between two adjacent lines of dialog as the sub-scene corresponding to the scene direction.

Based on these structures holding among image frames and sub-scene, we can retrieve an image frame that corresponds to a query, by the following procedure.

1. Input a query that consists of the time, location, player's name and her/his action.
2. Search the sub-scene that matches the condition stated in the input query using the existence/action map.
3. Extract a set of frame which correspond to the searched sub-scene, and display them on the image screen of user's GUI.

Our system uses Netscape Navigator as a GUI. The retrieval system is implemented as JAVA applets which work as a CGI. The following figures are screen images of GUI of our scene retrieval system.

Figure 2 is an introductory screen of GUI of our scene retrieval system. In it, introductory scenes of drama videos are displayed in every track.

Figure 3 is a screen in which several input forms are shown. Dousanushi('agent' in English) input form indicates a list of the players' name, from which we select one of the names. Then scenes in which the player of the selected name exists is searched. Basho(meaning 'location') and Jikoku(meaning 'time') input forms indicate the list of locations and the list of times, from which we select one value for each as query terms. Dousa(meaning 'action') input form indicates the list of verbs. If we select one of them, it comes to be a

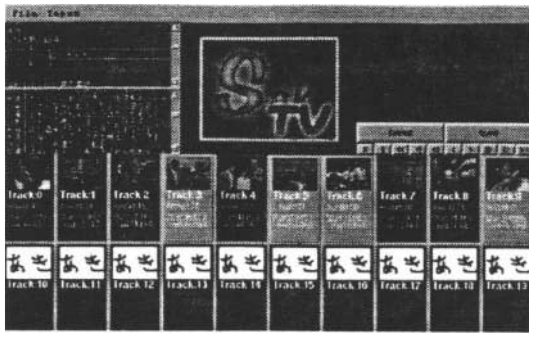


Figure 2: Introductory screen of GUI

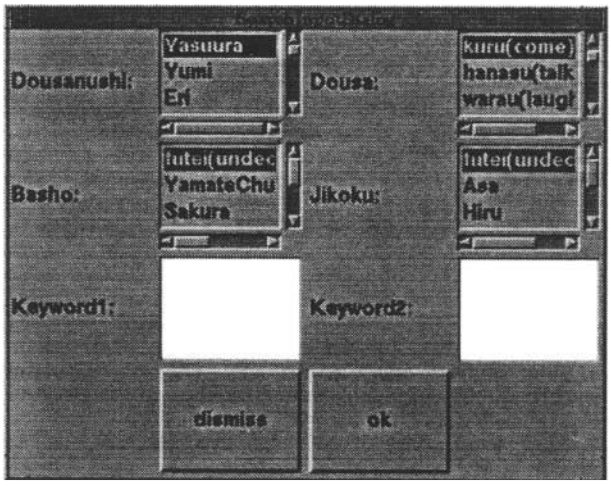


Figure 3: Forms for query input

term of the whole query, and specifys a verb appearing in the existence/action map. In Keyword1 and Keyword2 input forms we can write other keywords of retrieval condition. All these inputs are combined together to be used as one query. Then the retrieval system seeks scenes that meet all these conditions in the query by consulting the existence/action map.

Figure 4 is the result of retrieval. In the upper area, the contents of the query are shown. In the middle area, retrieved scenes are displayed. In the bottom area, track number that corresponds to the retrived drama scenes is shown. In this example, the query is as follows: player's name is "Yasuura", the action is "kuru('come')", the location and time are not specified, and no keywords are given. Then the player whose role name is Yasuura appears in all the retrieved scenes, and he is surely approaching to the camera in all of the retrieved scenes. Namely the system successfully retrieves the scenes that meet the query.

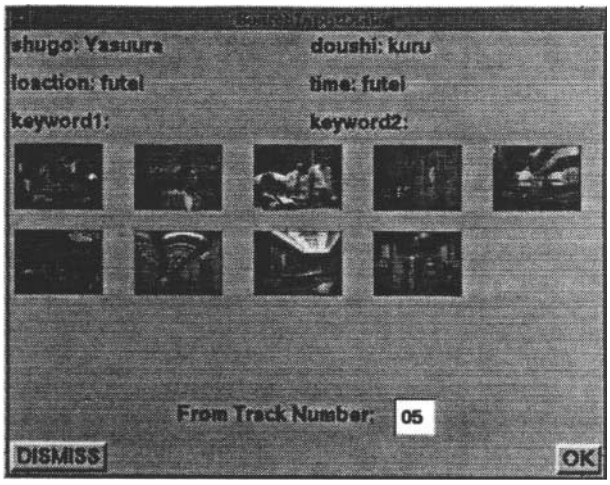


Figure 4: The retrieved scenes by the query input of Figure 3

5 Concluding Remarks

In this paper, we analyze sentences used as the scene direction, and classify them. Then we propose the rules to extract subjects and predicates from them. With the extracted subjects and predicates, we build the existence/action map that shows the situations happening on each scene. The existence/action map we build describes scenes very correctly, namely the recall is around 80% and the precision is over 90%. This means that our system of inferring reference relations works well for scene directions. Then we develop the scene retrieval system in which this map is used to retrieve scenes from the input query. We also show some experimental results of retrieval.

Our system is based on very simple linguistic rules. Therefore we expect that it is possible to improve the quality of retrieval by hiring more rules. However more sophisticated linguistic rules and processing mechanism are the open problem.

6 Acknowledgements

This work is supported in part by the Grant-in-Aid for Creative Basic Research: Multi Media Mediation Mechanism, of the Ministry of Education, Science, Sports and Culture, Japan.

References

Susan Brennan, Marilyn Walker Friedman and Carl Pollard. 1987. A Centering Approach to Pronouns. In *Proceedings of 25th Annual Meeting of ACL*. pages 155–162.

IPA. 1992. Dictionary for Computer of Basic Japanese Verbs(IPAL). Japan Information Promotion Association, Tokyo.

Megumi Kameyama. 1988. Zero Pronominal Binding: Where Syntax and Discourse Meet. In W. Poser, editors,*Japanese Syntax*, pages 47–73, CSLI, Stanford, CA.

Susumu Kuno. 1978. Danwa no Punpoo (Syntax of Discourse). Taishuukan, Tokyo.

Y. Matsumoto and S. Kurohashi and T. Utsuro and H. Taeki and M. Nagao. 1992. User's Manual of Japanese Morphological Analyzer: JUMAN version 2.0. Kyoto University, Kyoto.

Marilyn Walker and Masayo Iida and Sharon Cote. 1994. Japanese Discourse and the Process of Centering. *Computational Linguistics*, **20-2**, pages 193–232.

Yoshitomo Yaginuma and Masao Sakauchi. 1993. Moving TV Image Analysis Based on Multimedia Fusion Focusing on Extracted Common Concepts. *The proceedings of IEEE International Conference on Industrial Electronics'93*, IEEE.