Structural Analysis of Instruction Utterances using Linguistic and Visual Information

Tomohide Shibata, Masato Tachiki, Daisuke Kawahara,
Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida

Graduate School of Information Science and Technology, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{shibata, tachiki, kawahara, okamoto, kuro, nishida}@kc.t.u-tokyo.ac.jp

Abstract. In realizing video retrieval system, the crucial point is how to provide an effective access method of video contents. This paper focuses on Japanese cooking instruction utterances and describes a method of analyzing structure of them, which leads to a summary of video. We detect a hierarchical structure of video contents by using linguistic and visual information. We found that the integration of visual information can improve the detection of task units better than using linguistic information alone.

1 Introduction

The advance of computers, networks, and media processing techniques has made it possible to help human with intelligent works. For example, Internet search engines enable us to get text-based information immediately.

As an extension of such text-based systems, we can design a video-based system that teaches methods, notes, and tips interactively about various works. Such a system is effective in many fields, such as handiwork, cooking, sports, and so on, where ideally, we want to ask experts/teachers and follow their examples.

A simple way to realize such a system is to match a user's query with instruction utterances, such as speech recognition results and closed captions, and then, present the corresponding part of a video. In this case, there are two problems as follows:

- In instruction utterances, zero pronouns are often used, and there are not only explanations of actions but also tips of actions, notes, etc. Therefore, it is difficult to match a user's query with instruction utterances correctly.
- Since videos are time-sequential and redundant, it is very hard to skim videos as they are.

To solve these problems, we need organize instruction utterances. In the case of instruction video, we can exploit the structural information of instruction utterances of experts/teachers. This paper focuses on Japanese cooking instruction, and describes a method of analyzing structure of cooking instruction utterances. We do not deal with speech data but start with closed captions of TV cooking programs as shown in Figure 1.

```
Hello.
Welcome to "Today's Cooking".
Today, I'd like to introduce a suitable dish for this upcoming summer.
...
Next, cucumber.
I will put this in a soup.
Peel it.
Peel it with this peeler.
As you know that we can easily find cucumbers during summer,
and I think it will taste better if we add it in a soup.
Cut it lengthwise.
Then, cut it to diamonds.
```

Fig. 1. An example of cooking instruction utterances (NHK "Today's Cooking").

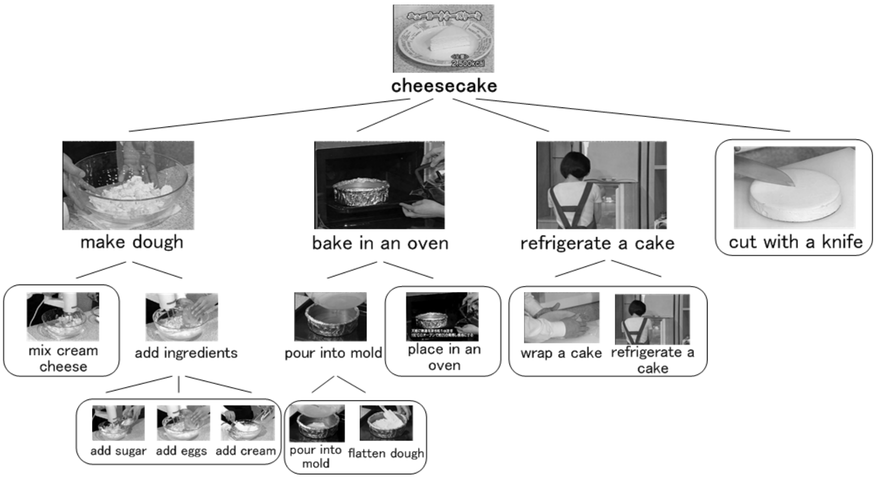


Fig. 2. Hierarchical structure in cooking.

Instruction utterances in cooking has a hierarchical tree structure. For example, a procedure of "making cheesecake" consists of making dough, baking a cake in an oven, refrigerating a cake, and so on. Making dough consists of mixing cream cheese and adding ingredients, as shown in Figure 2.

If such a hierarchical structure of video contents can be detected automatically, it can be utilized to make the summary of the video, which leads to a very effective access method to the video contents. In the next section, we explain how to analyze the instruction utterances based on linguistic information [1], and Section 3 explains the integration of visual information to the analysis.

2 Linguistic Analysis of Cooking Instruction Utterances

2.1 Anaphora Resolution

In Japanese, zero pronouns are often used, especially, in spoken language. Anaphora resolution of zero pronoun is inevitable to detect the discourse structure of utterances properly. In order to analyze anaphora, we have constructed a case frame dictionary in cooking domain. We built a zero pronoun resolution system using the case frame dictionary and the distance tendency that a zero pronoun has its antecedent in its close position [2].

Table 1. Utterance-types.

Utterance-type	Example
Action declaration:	Then, we cook a steak.
Individual action:	Pour water into a pan.
Food state:	There is no water in a carrot.
Food/Tool presentation:	Ingredients are 150g minced beef.
Substitution:	You may squeeze it by hand.
Note:	Be careful so that seeds won't go into.
Miscellaneous:	Hello.

Table 2. Examples of patterns for attaching utterance-type.

Pattern	Example
action declaration	
··· *ni-kakarimasu*(begin to ···)	Then, I begin to cook a steak.
··· *te-ikimasu*(be going to ···)	And then, I will add spice.
food state	
intransitive verb	Water boiled.
adjective + *naru*(become)	Oil heated up sufficiently.
food/tool presentation	
<food/tool> *wo-tsukaimasu*(use)	I use this handy mixer.
substitution	
··· *shitemo-kekkoudesu*(may)	You may use lemon juice.
··· *demo-kekkoudesu*(may)	You may use sliced meat.

2.2 Utterance-Type Detection

In cooking instruction utterances, while explanations of actions are dominant, there are several types of utterances such as declaration of beginning of series of actions, tips of actions, notes, etc. among them. We classify cooking instruction utterance into 7 types as shown in Table 1, by referring to Izuno [3].

Among them, action declaration, food/tool presentation, substitution, notes, and miscellaneous can be recognized by patterns of sentence-end. As for individual action and food state, we use general rules regarding intransitive verbs or adjective + *"naru"*(become) as food state, and others as individual action. Examples of patterns for attaching utterance-type are shown in Table 2.

2.3 Discourse Structure Analysis

Based on types attached in the previous section, we analyze discourse structure of utterances. As a model of the discourse structure, we suppose a graph structure that each utterance is one node and is linked with related utterances. In the discourse structure of task-oriented utterance like cooking, a task reflecting the tree structure is the core, and such utterance as substitution and notes modify it.

A method of the discourse structure analysis is based on [4]. As a new sentence comes in, by checking surface information, we find a connected sentence and the coherence relation between them. Examples of rules for discourse structure analysis are shown in Table 3. Each rule specifies a condition for a pair of a new sentence and a possible connected sentence: the range of possible connected

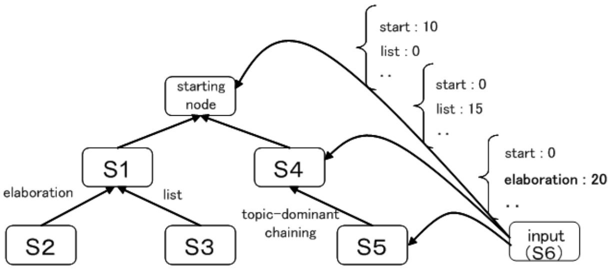


Fig. 3. A model of discourse structure.

Table 3. Examples of rules of discourse structure analysis.

Coherence relation	Score	Applicable range	Patterns for a connected sentence	Patterns for a new sentence
list	5	1	*	*soshite*(then)· · ·
contrast	30	1	*	*mushiro*(rather than)· · ·
contrast	40	*	X· · ·	X'($\simeq$ X)· · ·
elaboration	15	1	*	<note>
reason	30	1	*	· · ·karada(because)

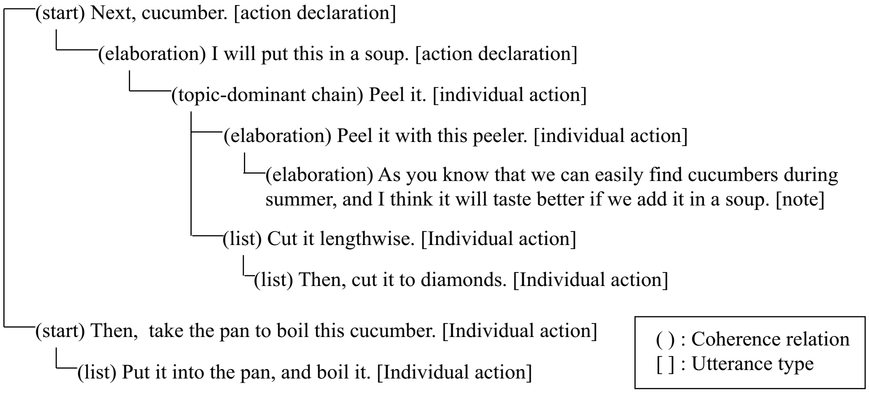
sentences (how far from the new sentence) and patterns for the two sentences. If a pair meets these condition, the relation and score in the rule are given to it. As a final result, we choose the connected sentence and the relation that have the maximum score. Figure 4 shows the discourse structure of utterances in Figure 1.

Once we can detect the discourse structure of utterances, we can utilize it to make a summary of the video. A straightforward way is to segment the video at each "start" relation point. A resultant segment can be considered as action sequence, which we call *task unit* hereafter (In Figure 2, each box denotes the task unit). By detecting a representative frame and utterance of each task unit (possibly the first frame and the first action utterance in the unit), we can make a summary of the video.

3 Integration of Linguistic and Visual Information

It is still very difficult to analyze spoken language accurately, detecting appropriate relations between utterances. The discourse analysis method explained in the previous section tends to make smaller segments because of the failure of detecting relations between utterances. This section proposes a method of integrating linguistic and visual information to improve the accuracy of analyzing discourse structure.

A video clip consists of a sequence of frames. A shot is defined as a collection of frames captured from a single camera operation, and a cut point is an instantaneous change from one shot to another (we call the first frame after the cut point a cut frame). A scene is defined as a sequential collection of shots unified by a common event. In the case of instruction video, face shots and hand shots come one after another, and one or more pairs of face shot and hand shot compose a scene, which corresponds to a task unit introduced in the previous



(start) Next, cucumber. [action declaration]

(elaboration) I will put this in a soup. [action declaration]

(topic-dominant chain) Peel it. [individual action]

(elaboration) Peel it with this peeler. [individual action]

(elaboration) As you know that we can easily find cucumbers during summer, and I think it will taste better if we add it in a soup. [note]

(list) Cut it lengthwise. [Individual action]

(list) Then, cut it to diamonds. [Individual action]

(start) Then, take the pan to boil this cucumber. [Individual action]

(list) Put it into the pan, and boil it. [Individual action]

( ) : Coherence relation
[ ] : Utterance type

Fig. 4. An example of discourse structure.

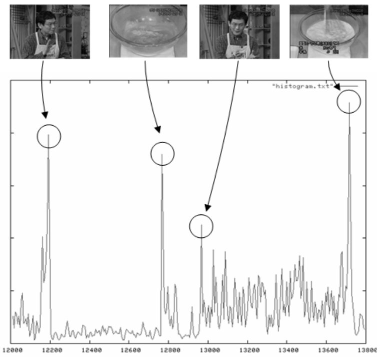


Fig. 5. An example of cut frame detection.

section. During the first face shot of a task unit, an instructor explains an action which he/she will perform.

Since the instruction video has such a visual structure, we can integrate visual information with linguistic information to improve the detection of task unit. Roughly speaking, if an utterance is a boundary both linguistically (start relation is detected) and visually (cut frame of face shot), it can be considered as the starting utterance of a new task unit.

Various automatic cut point detection algorithms have been proposed [5]. We employ the method of using color histogram difference. If color histogram difference exceeds a threshold, we regard this point as a cut point (Figure 5). This method is relatively simple, but achieves high accuracy. We also implemented a classifier of face cut frame and hand cut frame using facial templates.

After detecting face cut frames, we try to classify them into the following three classes:

(a) the utterance at the cut frame is a start of a new task unit.
(b) the next utterance is a start of a new task unit.
(c) the cut does not correspond to a task unit boundary.

A classifier of face cut frames were constructed based on the following feature set.

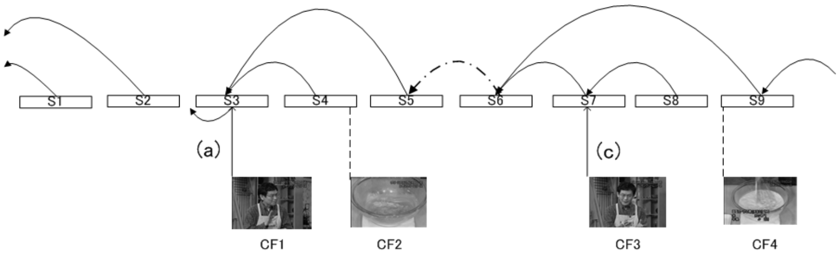


Fig. 6. Classification of face cut frames and modification of discourse structure.

- Difference value of the cut point in the color histogram.
- Utterance-types of the previous, corresponding, and next utterances.
- Coherence relation of the corresponding utterance.
- Cue phrases such as "then", "now", etc. in the corresponding utterance.
- Time duration from the previous cut and that to the next cut.
- Silence duration within 3 seconds window.

We use Support Vector Machines(SVM). Since SVM is a binary classifier, we adopt the pairwise method, which constructs classifiers for each pair of the above three classes, (a), (b) and (c).

Figure 6 shows an example. Suppose the two face cut frames were detected, and the cut face frame 1 (CF1) was classified to (a) and the cut face frame 3 (CF3) to (c). These classes are consistent with the discourse structure (the result of linguistic analysis). On the other hand, though the sentence 6 (S6) was detected as a task unit boundary based on the discourse structure analysis, no face cut frame was detected here, and the discourse structure is modified to make S6 depend on S5.

4 Experimental Results

We used the corpus of NHK TV program, "Today's Cooking". A program consists of about 200 utterances, and the average length of an utterance is about 20 characters. Five programs were used for the experiments. In these programs, 150 face cut frames were detected automatically, when the threshold of the difference of color histogram was set to 15,000,000. Then, we labeled them (a), (b), or (c) manually. The resultant data was used as training/testing data for 5-hold cross validation of SVM classifier. We employed TinySVM[1] with polynomial kernels of degree 2.

The result of face cut frame classification is shown in Table 4, and the result of task unit detection is shown is shown in Table 5. The accuracy of face cut frame classification was 74%, and that information can be utilized to improve the task unit detection with linguistic information.

5 Conclusion and Future Work

In this paper, we described a method of analyzing structure of cooking instruction utterances. We found that the integration of visual information can improve the detection of task units better than using linguistic information alone.

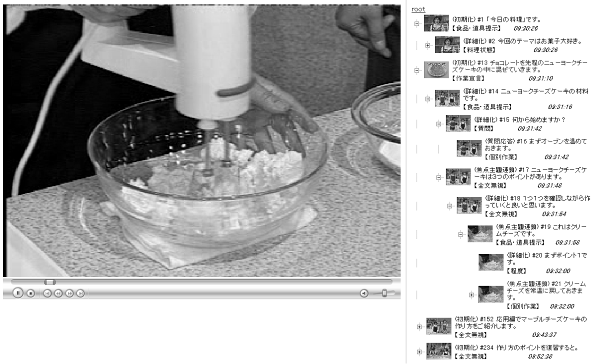
[1] http://cl-aist-nara.ac.jp/˜taku-ku/software/TinySVM/

Table 4. Experimental results of face cut frame classification.

# of data	100	120	150
Closed	93.0%	92.5%	89.3%
Open	65.0%	72.5%	74.0%
Baseline	63.0%	65.8%	65.3%

Table 5. Experimental results of task unit detection.

	Linguistic info.	Linguistic & visual info.
Precision	28.0%(50/176)	45.8%(33/72)
Recall	62.5%(50/80)	41.2%(33/80)
F	0.392	**0.434**



root

(初期化) #1「今日の料理」です。
【食品・道具提示】 09:30:26

(詳細化) #2 今回のテーマはお菓子大好き。
【料理状態】 09:30:26

(初期化) #13 チョコレートを先程のニューヨークチーズケーキの中に混ぜていきます。
【作業宣言】 09:31:10

(詳細化) #14 ニューヨークチーズケーキの材料です。
【食品・道具提示】 09:31:16

(詳細化) #15 何から始めますか？
【質問】 09:31:42

(質問応答) #16 まずオーブンを温めておきます。
【個別作業】 09:31:42

(焦点主題連鎖) #17 ニューヨークチーズケーキは3つのポイントがあります。
【全文無視】 09:31:48

(詳細化) #18 1つ1つを確認しながら作っていくと良いと思います。
【全文無視】 09:31:54

(焦点主題連鎖) #19 これはクリームチーズです。
【食品・道具提示】 09:31:58

(詳細化) #20 まずポイント1です。
【程度】 09:32:00

(焦点主題連鎖) #21 クリームチーズを常温に戻しておきます。
【個別作業】 09:32:00

(初期化) #152 応用編でマーブルチーズケーキの作り方をご紹介します。
【全文無視】 09:43:37

(初期化) #234 作り方のポイントを復習すると。
【全文無視】 09:52:38

Fig. 7. A video archive retrieval system.

We are constructing video retrieval system of cooking instruction video archive (Figure 7) and planning to make a summary based on the detected boundaries.

References

1. Tomohide Shibata, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida. Structural analysis of instruction utterances. In *Proceedings of Seventh International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2003)*, pp. 1054–1061, September 2003.
2. Daisuke Kawahara and Sadao Kurohashi. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of The 1st International Joint Conference on Natural Language Processing*, pp. 334–341, 2004.
3. Hidekatsu Izuno, Yuichi Nakamura, and Yuichi Ohta. Quevico: A framework for video-based interactive media. In *Working Notes WS-5 International Workshop on Intelligent Media Technology for Communicative Reality, PRICAI-02 (Seventh Pacific Rim International Conference on Artificial Intelligence)*, pp. 6–11, August 2002.
4. Sadao Kurohashi and Makoto Nagao. Automatic detection of discourse structure by checking surface information in sentences. In *Proceedings of 15th COLING*, Vol. 2, pp. 1123–1127, 1994.
5. Rainer Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proceedings of SPIE Conf. on Storage and Retrieval for Image & Video Databases VII*, Vol. 3656, pp. 290–301, 1998.