Unsupervised Topic Identification by Integrating Linguistic and Visual Information Based on Hidden Markov Models

Tomohide Shibata
Graduate School of Information Science
and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656, Japan
shibata@kc.t.u-tokyo.ac.jp

Sadao Kurohashi
Graduate School of Informatics,
Kyoto University
Yoshida-honmachi, Sakyo-ku,
Kyoto, 606-8501, Japan
kuro@i.kyoto-u.ac.jp

Abstract

This paper presents an unsupervised topic identification method integrating linguistic and visual information based on Hidden Markov Models (HMMs). We employ HMMs for topic identification, wherein a state corresponds to a topic and various features including linguistic, visual and audio information are observed. Our experiments on two kinds of cooking TV programs show the effectiveness of our proposed method.

1 Introduction

Recent years have seen the rapid increase of multimedia contents with the continuing advance of information technology. To make the best use of multimedia contents, it is necessary to segment them into meaningful segments and annotate them. Because manual annotation is extremely expensive and time consuming, automatic annotation technique is required.

In the field of video analysis, there have been a number of studies on shot analysis for video retrieval or summarization (highlight extraction) using Hidden Markov Models (HMMs) (e.g., (Chang et al., 2002; Nguyen et al., 2005; Q.Phung et al., 2005)). These studies first segmented videos into shots, within which the camera motion is continuous, and extracted features such as color histograms and motion vectors. Then, they classified the shots based on HMMs into several classes (for baseball sports video, for example, pitch view, running overview or audience view). In these studies, to achieve high accuracy, they relied on handmade domain-specific knowledge or trained HMMs with manually labeled data. Therefore, they cannot be easily extended to new domains on a large scale. In addition, although linguistic information, such as narration, speech of characters, and commentary, is intuitively useful for shot analysis, it is not utilized by many of the previous studies. Although some studies attempted to utilize linguistic information (Jasinschi et al., 2001; Babaguchi and Nitta, 2003), it was just keywords.

In the field of Natural Language Processing, Barzilay and Lee have recently proposed a probabilistic content model for representing topics and topic shifts (Barzilay and Lee, 2004). This content model is based on HMMs wherein a state corresponds to a topic and generates sentences relevant to that topic according to a state-specific language model, which are learned from raw texts via analysis of word distribution patterns.

In this paper, we describe an unsupervised topic identification method integrating linguistic and visual information using HMMs. Among several types of videos, in which instruction videos (*how-to* videos) about sports, cooking, D.I.Y., and others are the most valuable, we focus on cooking TV programs. In an example shown in Figure 1, *preparation*, *sauteing*, and *dishing up* are automatically labeled in sequence. Identified topics lead to video segmentation and can be utilized for video summarization.

Inspired by Barzilay's work, we employ HMMs for topic identification, wherein a state corresponds to a topic, like *preparation* and *frying*, and various features, which include visual and audio information as well as linguistic information (instructor's utterances), are observed. This study considers a clause as an unit of analysis and the following eight topics as a set of states: *preparation, sauteing, frying, baking, simmering, boiling, dishing up, steaming*.

In Barzilay's model, although domain-specific

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 755–762,
Sydney, July 2006.

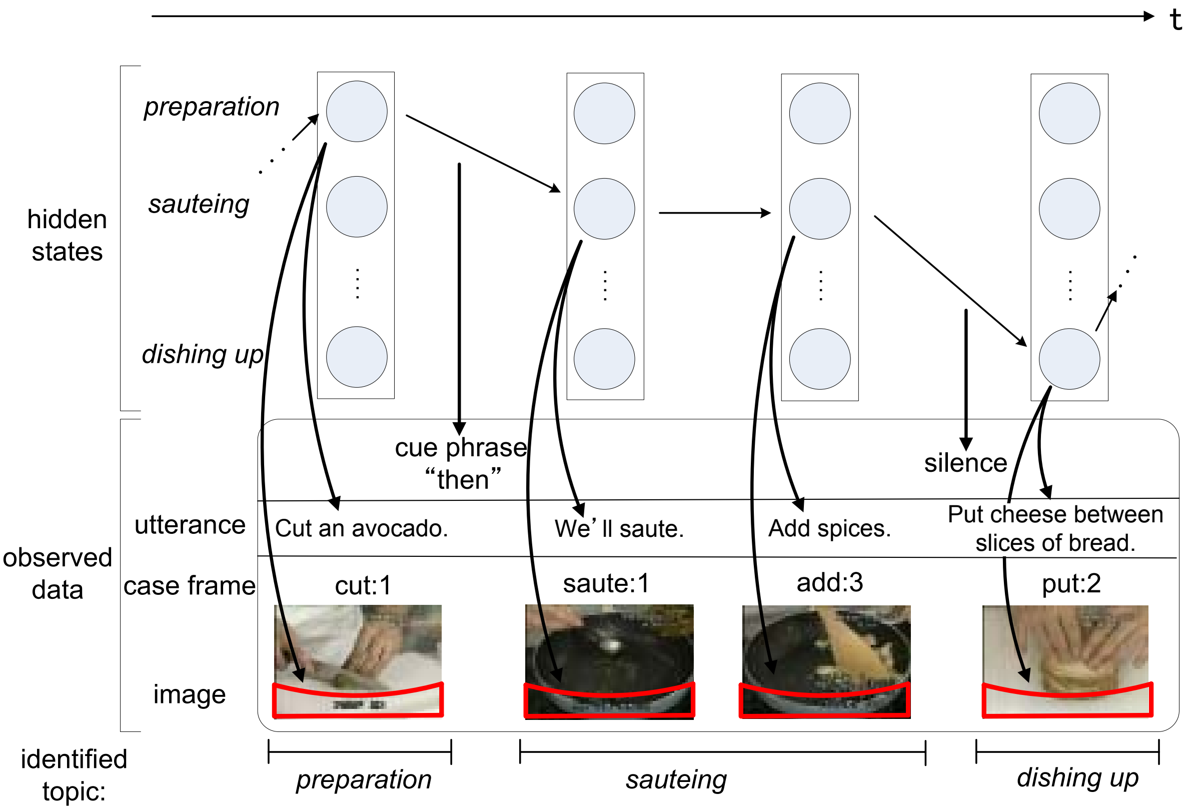


Figure 1: Topic identification with Hidden Markov Models.

word distribution can be learned from raw texts, their model cannot utilize discourse features, such as cue phrases and lexical chains. We incorporate domain-independent discourse features such as cue phrases, noun/verb chaining, which indicate topic change/persistence, into the domain-specific word distribution.

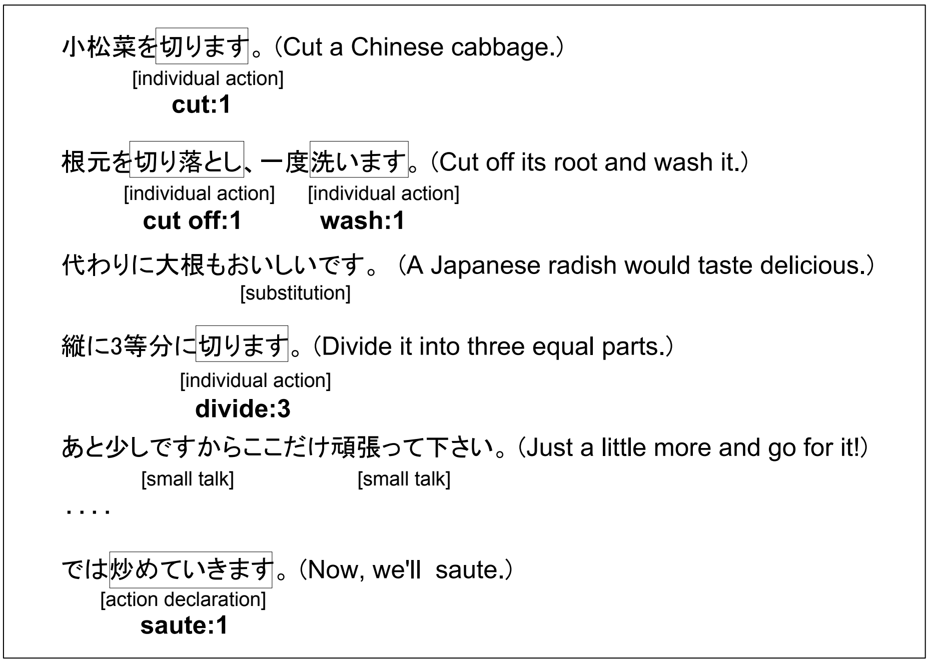
Our main claim is that we utilize visual and audio information to achieve robust topic identification. As for visual information, we can utilize background color distribution of the image. For example, *frying* and *boiling* are usually performed on a gas range and *preparation* and *dishing up* are usually performed on a cutting board. This information can be an aid to topic identification. As for audio information, silence can be utilized as a clue to a topic shift.

2 Related Work

In Natural Language Processing, text segmentation tasks have been actively studied for information retrieval and summarization. Hearst proposed a technique called TextTiling for subdividing texts into sub-topics (Hearst.M, 1997). This method is based on lexical co-occurrence. Galley et al. presented a domain-independent topic segmentation algorithm for multi-party speech (Galley et al., 2003). This segmentation algorithm uses automatically induced decision rules to combine linguistic features (lexical cohesion and cue phrases) and speech features (silences, overlaps and speaker change). These studies aim just at segmenting a given text, not at identifying topics of segmented texts.

Marcu performed rhetorical parsing in the framework of Rhetorical Structure Theory (RST) based on a discourse-annotated corpus (Marcu, 2000). Although this model is suitable for analyzing local modification in a text, it is difficult for this model to capture the structure of topic transition in the whole text.

In contrast, Barzilay and Lee modeled a content structure of texts within specific domains, such as earthquake and finance (Barzilay and Lee, 2004). They used HMMs wherein each state corresponds to a distinct topic (e.g., in earthquake domain, earthquake magnitude or previous earthquake occurrences) and generates sentences relevant to that topic according to a state-specific language model. Their method first create clusters via complete-link clustering, measuring sentence similarity by the cosine metric using word bigrams as features. They calculate initial probabilities: state s_i specific language model $p_{s_i}(w'|w)$



小松菜を切ります。（Cut a Chinese cabbage.）
[individual action]
cut:1

根元を切り落とし、一度洗います。（Cut off its root and wash it.）
[individual action] [individual action]
cut off:1 wash:1

代わりに大根もおいしいです。（A Japanese radish would taste delicious.）
[substitution]

縦に3等分に切ります。（Divide it into three equal parts.）
[individual action]
divide:3

あと少しですからここだけ頑張って下さい。（Just a little more and go for it!）
[small talk] [small talk]

. . . .

では炒めていきます。（Now, we'll saute.）
[action declaration]
saute:1

Figure 2: An example of closed captions. (The phrase sandwiched by a square bracket means an utterance type and the word surrounded by a rectangle means an extracted utterance referring to an action. The bold word means a case frame assigned to the verb.)

and state-transition probability $p(s_j|s_i)$ from state s_i to state s_j. Then, they continue to estimate HMM parameters with the Viterbi algorithm until the clustering stabilizes. They applied the constructed content model to two tasks: information ordering and summarization. We differ from this study in that we utilize multimodal features and domain-independent discourse features to achieve robust topic identification.

In the field of video analysis, there have been a number of studies on shot analysis with HMMs. Chang et al. described a method for classifying shots into several classes for highlight extraction in baseball games (Chang et al., 2002). Nguyen et al. proposed a robust statistical framework to extract highlights from a baseball video (Nguyen et al., 2005). They applied multi-stream HMMs to control the weight among different features, such as principal component features capturing color information and frame-difference features for moving objects. Phung et al. proposed a probabilistic framework to exploit hierarchy structure for topic transition detection in educational videos (Q.Phung et al., 2005).

Some studies attempted to utilize linguistic information in shot analysis (Jasinschi et al., 2001; Babaguchi and Nitta, 2003). For example, Babaguchi and Nitta segmented closed caption text into meaningful units and linked them to video streams in sports video. However, linguistic information they utilized was just keywords.

3 Features for Topic Identification

First, we'll describe the features that we use for topic identification, which are listed in Table 1. They consist of three modalities: linguistic, visual and audio modality.

We utilize as linguistic information the instructor's utterances in video, which can be divided into various types such as actions, tips, and even small talk. Among them, actions, such as cut, peel and grease a pan, are dominant and supposed to be useful for topic identification and others can be noise.

In the case of analyzing utterances in video, it is natural to utilize visual information as well as linguistic information for robust analysis. We utilize background image as visual information. For example, *frying* and *boiling* are usually performed on a gas range and *preparation* and *dishing up* are usually performed on a cutting board.

Furthermore, we utilize cue phrases and silence as a clue to a topic shift, and noun/verb chaining as a clue to a topic persistence.

We describe these features in detail in the following sections.

3.1 Linguistic Features

Closed captions of Japanese cooking TV programs are used as a source for extracting linguistic fea-

Table 1: Features for topic identification.

Modality	Feature	Domain dependent	Domain independent
linguistic	case frame	utterance generalization	
	cue phrases		topic change
	noun chaining		topic persistence
	verb chaining		topic persistence
visual	background image	bottom of image	
audio	silence		topic change

Table 2: Utterance-type classification. (An underlined phrase means a pattern for recognizing utterance type.)

[action declaration]
ex. さ，では，ステーキにかかります．　(Then, we ’ll cook a steak)
じゃあ炒めていきましょう．　(OK, we’ll fry.)
[individual action]
ex. なすはヘタを取ります。 (Cut off a step of this eggplant.)
お鍋にお水を入れます．　(Pour water into a pan.)
[food state]
ex. ニンジンの水分がなくなりました．　(There is no water in the carrot.)
[note]
ex. 芯は切らないで下さい．　(Don’t cut this core off.)
[substitution]
ex. 青ねぎでも結構です．　(You may use a leek.)
[food/tool presentation]
ex. 今日はこのハンドミキサーを使います．　Today, we use this handy mixer.)
[small talk]
ex. こんにちは．　(Hello.)

tures. An example of closed captions is shown in Figure 2. We first process them with the Japanese morphological analyzer, JUMAN (Kurohashi et al., 1994), and make syntactic/case analysis and anaphora resolution with the Japanese analyzer, KNP (Kurohashi and Nagao, 1994). Then, we perform the following process to extract linguistic features.

3.1.1 Extracting Utterances Referring to Actions

Considering a clause as a basic unit, utterances referring to an action are extracted in the form of case frame, which is assigned by case analysis. This procedure is performed for generalization and word sense disambiguation. For example, “塩を入れる (add salt)” and “砂糖を鍋に入れる (add sugar into a pan)” are assigned to case frame ireru:1 (add) and “包丁を入れる (carve with a knife)” is assigned to case frame ireru:2 (carve). We describe this procedure in detail below.

Utterance-type recognition

To extract utterances referring to actions, we classify utterances into several types listed in Table 2[1]. Note that actions are supposed to have two levels: [action declaration] means a declaration of beginning a series of actions and [individual action] means an action that is the finest one.

Input sentences are first segmented into clauses and their utterance type is recognized. Among several utterance types, [individual action], [food/tool presentation], [substitution], [note], and [small talk] can be recognized by clause-end patterns. We prepare approximately 500 patterns for recognizing the utterance type. As for [individual action] and [food state], considering the portability of our system, we use general rules regarding intransitive verbs or adjective + “なる (become)” as [food state], and others as [individual action].

Action extraction

We extract utterances whose utterance type is recognized as action ([action declaration] or [individual action]). For example, “むく (peel)” and “切る (cut)” are extracted from the following sentence.

(1) にんじんは皮をむき [individual action] 半分の長さに切ります [individual action]。(We peel this carrot and cut it in half.)

We make two exceptions to reduce noises. One is that clauses are not extracted from the sentence in which sentence-end clause’s utterance-type is not recognized as an action. In the following example, “煮る (simmer)” and “切る (cut)” are not extracted because the utterance type of

[1] In this paper, [ ] means an utterance type.

Table 3: An example of the automatically constructed case frame.

Verb	Case marker	Examples
kiru:1	*ga*	<agent>
(cut)	*wo*	pork, carrot, vegetable, · · ·
	ni	rectangle, diamonds, · · ·
kiru:2	*ga*	<agent>
(drain)	*wo*	damp · · ·
	no	eggplant, bean curd, · · ·
ireru:1	*ga*	<agent>
(add)	*wo*	salt, oil, vegetable, · · ·
	ni	pan, bowl, · · ·
ireru:2	*ga*	<agent>
(carve)	*wo*	knife · · ·
	ni	fish · · ·

the sentence-end clause is recognized as [substitution].

(2) 煮てから [individual action] 切っても [individual action] 構いません [substitution]。(It doesn't matter if you cut it after simmering.)

The other is that conditional/causal clauses are not extracted because they sometimes refer to the previous/next topic.

(3) 切りましたら 炒めていきます。(After we finish cutting it, we'll fry.)

(4) プチトマトは油で 揚げるので、切り込みを入れます。(We cut in this cherry tomato, because we'll fry it in oil.)

Note that relations between clauses are recognized by clause-end patterns.

Verb sense disambiguation by assigning to a case frame

In general, a verb has multiple meanings/usages. For example, "入れる" has multiple usages, "塩を入れる (add salt)" and "包丁を入れる (carve with a knife)" , which appear in different topics. We do not extract a surface form of verb but a case frame, which is assigned by case analysis. Case frames are automatically constructed from Web cooking texts (12 million sentences) by clustering similar verb usages (Kawahara and Kurohashi, 2002). An example of the automatically constructed case frame is shown in Table 3. For example, "塩を入れる (add salt)" is assigned to ireru:1 (add) and "包丁を入れる (carve with a knife)" is assigned to case frame ireru:2 (carve).

3.1.2 Cue phrases

As Grosz and Sidner (Grosz and Sidner, 1986) pointed out, cue phrases such as *now* and *well* serve to indicate a topic change. We use approximately 20 domain-independent cue phrases, such as "では (then)", "次は (next)" and "そうしましたら (then)".

3.1.3 Noun Chaining

In text segmentation algorithms such as TextTiling (Hearst.M, 1997), lexical chains are widely utilized for detecting a topic shift. We utilize such a feature as a clue to topic persistence.

When two continuous actions are performed to the same ingredient, their topics are often identical. For example, because "おろす (grate)" and "上げる (raise)" are performed to the same ingredient "かぶら (turnip)" , the topics (in this instance, *preparation*) in the two utterances are identical.

(5) a. かぶらをおろし金でおろしていきます。(We'll grate a turnip.)
b. おろしたかぶらをざるに上げます。(Raise this turnip on this basket.)

However, in the case of spoken language, because there exist many omissions, it is often the case that noun chaining cannot be detected with surface word matching. Therefore, we detect noun chaining by using the anaphora resolution result[2] of verbs (ex.(6)) and nouns (ex.(7)). The verb, noun anaphora resolution is conducted by the method proposed by (Kawahara and Kurohashi, 2004), (Sasano et al., 2004), respectively.

(6) a. キャベツを切ります。 (Cut a cabbage.)
b. 一度 [キャベツを] 洗います。 (Wash it once.)

(7) a. にんじんを大体４ｃｍくらい切ります。(Slice a carrot into 4-cm pieces.)
b. [にんじんの] 皮をぐるっとむきます。(Peel its skin.)

3.1.4 Verb Chaining

When a verb of a clause is identical with that of the previous clause, they are likely to have the same topic. We utilize the fact that the adjoining two clauses contain an identical verbs or not as an observed feature.

(8) a. とうがらしを入れて下さい。(Add some red peppers.)

[2] [ ] indicates an element complemented with anaphora resolution.

b. 鶏手羽を入れます。 (Add chicken wings.)

3.2 Image Features

It is difficult for the current image processing technique to extract what object appears or what action is performing in video unless a detailed object/action model for a specific domain is constructed by hand. Therefore, referring to (Hamada et al., 2000), we focus our attention on color distribution at the bottom of the image, which is comparatively easy to exploit. As shown in Figure 1, we utilize the mass point of RGB in the bottom of the image at each clause.

3.3 Audio Features

A cooking video contains various types of audio information, such as instructor's speech, cutting sounds and frizzling sound. If cutting sound or frizzling sound could be distinguished from other sounds, they could be an aid to topic identification, but it is difficult to recognize them.

As Galley et al. (Galley et al., 2003) pointed out, a longer silence often appears when topic changes, and we can utilize it as a clue to topic change. In this study, silence is automatically extracted by finding duration below a certain amplitude level which lasts more than one second.

4 Topic Identification based on HMMs

We employ HMMs for topic identification, where a hidden state corresponds to a topic and various features described in Section 3 are observed. In our model, considering the case frame as a basic unit, the case frame and background image are observed from the state, and discourse features indicating to topic shift/persistence (cue phrases, noun/verb chaining and silence) are observed when the state transits.

4.1 Parameters

HMM parameters are as follows:

- initial state distribution π_i : the probability that state s_i is a start state.
- state transition probability a_{ij} : the probability that state s_i transits to state s_j.
- observation probability $b_{ij}(o_t)$: the probability that symbol o_t is emitted when state s_i transits to state s_j. This probability is given by the following equation:

$$b_{ij}(o_t) = b_j(cf_k) \cdot b_j(R,G,B) \cdot b_{ij}(discourse\ features) \quad (1)$$

- case frame $b_j(cf_k)$: the probability that case frame cf_k is emitted by state s_j.

- background image $b_j(R,G,B)$: the probability that background image $b_j(R,G,B)$ is emitted by state s_j. The emission probability is modeled by a single Gaussian distribution with mean (R_j,G_j,B_j) and variance σ_j.

- discourse features : the probability that discourse features are emitted when state s_i transits to state s_j. This probability is defined as multiplication of the observation probability of each feature (cue phrase, noun chaining, verb chaining, silence). The observation probability of each feature does not depend on state s_i and s_j, but on whether s_i and s_j are the same or different. For example, in the case of cue phrase (c), the probability is given by the following equation:

$$b_{ij}(c) = \begin{cases} p_{same}(c)(i = j) \\ p_{diff}(c)(i \neq j) \end{cases} \quad (2)$$

4.2 Parameters Estimation

We apply the Baum-Welch algorithm for estimating these parameters. To achieve high accuracy with the Baum-Welch algorithm, which is an unsupervised learning method, some labeled data have been required or proper initial parameters have been set depending on domain-specific knowledge. These requirements, however, make it difficult to extend to other domains. We automatically extract "pseudo-labeled" data focusing on the following linguistic expressions: if a clause has the utterance-type [action declaration] and an original form of its verb corresponds to a topic, its topic is set to that topic. Remind that [action declaration] is a kind of declaration of starting a series of actions. For example, in Figure 1, the topic of the clause "We'll saute." is set to *sauteing* because its utterance-type is recognized as [action declaration] and the original form of its verb is topic *sauteing*.

By using a small amounts of "pseudo-labeled" data as well as unlabeled data, we train the HMM parameters. Once the HMM parameters are trained, the topic identification is performed using the standard Viterbi algorithm.

5 Experiments and Discussion

5.1 Data

To demonstrate the effectiveness of our proposed method, we made experiments on two kinds of cooking TV programs: NHK "Today's Cooking"

Table 5: Experimental result of topic identification.

Features				Accuracy	
case frame	background image	discourse features	silence	“Today’s Cooking”	“Kewpie 3-Min Cooking”
√				61.7%	66.4%
	√			56.8%	72.9%
√	√			69.9%	77.1%
√	√	√		**70.5%**	**82.9%**
√	√	√	√	70.5%	82.9%

Table 4: Characteristics of the two cooking programs we used for our experiments.

Program	Today’s Cooking	Kewpie 3-Min Cooking
Videos	200	70
Duration	25min	10min
# of utterances per video	249.4	183.4

and NTV “Kewpie 3-Min Cooking”. Table 4 presents the characteristics of the two programs. Note that time stamps of closed captions synchronize themselves with the video stream. Extracted “pseudo-labeled” data by the expression mentioned in Section 4.2 are 525 clauses out of 13564 (3.87%) in “Today’s Cooking”, and 107 clauses out of 1865 (5.74%) in “Kewpie 3-Min Cooking”.

5.2 Experiments and Discussion

We conducted the experiment of the topic identification. We first trained HMM parameters for each program, and then applied the trained model to five videos each, in which, we manually assigned appropriate topics to clauses. Table 5 gives the evaluation results. The unit of evaluation was a clause. The accuracy was improved by integrating linguistic and visual information compared to using linguistic / visual information alone. (Note that “visual information” uses pseudo-labeled data.) In addition, the accuracy was improved by using various discourse features. The reason why silence did not contribute to accuracy improvement is supposed to be that closed captions and video streams were not synchronized precisely due to time lagging of closed captions. To deal with this problem, an automatic closed caption alignment technique (Huang et al., 2003) will be applied or automatic speech recognition will be used as texts instead of closed captions with the advance of speech recognition technology.

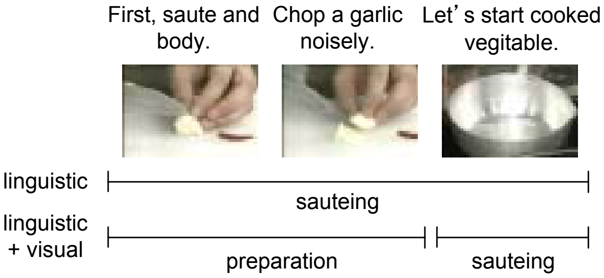


Figure 3: An improved example by adding visual information.

Figure 3 illustrates an improved example by adding visual information. In the case of using only linguistic information, this topic was recognized as *sauteing*, but this topic was actually *preparation*, which referred to the next topic. By using the visual information that background color was white, this topic was correctly recognized as *preparation*.

We conducted another experiment to demonstrate the validity of several linguistic processes, such as utterance-type recognition and word sense disambiguation with case frames, for extracting linguistic information from closed captions described in Section 3.1.1. We compared our method to three methods: a method that does not perform word sense disambiguation with case frames (**w/o cf**), a method that does not perform utterance-type recognition for extracting actions (uses all utterance-type texts) (**w/o utype**), a method, in which a sentence is emitted according to a state-specific language model (bigram) as Barzilay and Lee adopted (**bigram**). Figure 6 gives the experimental result, which demonstrates our method is appropriate.

One cause of errors in topic identification is that some case frames are incorrectly constructed. For example, *kiru*:1 (cut) contains “野菜を切る (cut a vegetable)” and “油を切る (drain oil)”. This leads to incorrect parameter training. Other cause is that some verbs are assigned to an inaccurate case frame by the failure of case analysis.

6 Conclusions

This paper has described an unsupervised topic identification method integrating linguistic and visual information based on Hidden Markov Mod-

Table 6: Results of the experiment that compares our method to three methods.

Method	Accuracy	
	“Today’s Cooking”	“Kewpie 3-Min Cooking”
proposed method	**61.7%**	**66.4%**
w/o cf	57.1%	60.0%
w/o utype	61.7%	62.1%
bigram	54.7%	59.3%

els. Our experiments on the two kinds of cooking TV programs showed the effectiveness of integration of linguistic and visual information and incorporation of domain-independent discourse features to domain-dependent features (case frame and background image).

We are planning to perform object recognition using the automatically-constructed object model and utilize the object recognition results as a feature for HMM-based topic identification.

References

Noboru Babaguchi and Naoko Nitta. 2003. Intermodal collaboration: A strategy for semantic content analysis for broadcasted sports video. In *Proceedings of IEEE International Conference on Image Processing(ICIP2003)*, pages 13–16.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the NAACL/HLT*, pages 113–120.

Peng Chang, Mei Han, and Yihong Gong. 2002. Extract highlights from baseball game video with hidden markov models. In *Proceedings of the International Conference on Image Processing 2002(ICIP2002)*, pages 609–612.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, 7.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistic*, 12:175–204.

Reiko Hamada, Ichiro Ide, Shuichi Sakai, and Hidehiko Tanaka. 2000. Associating cooking video with related textbook. In *Proceedings of ACM Multimedia 2000 workshops*, pages 237–241.

Hearst.M. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March.

Chih-Wei Huang, Winston Hsu, and Shin-Fu Chang. 2003. Automatic closed caption alignment based on speech recognition transcripts. Technical report, Columbia ADVENT.

Radu Jasinschi, Nevenka Dimitrova, Thomas McGee, Lalitha Agnihotri, John Zimmerman, and Dongge. 2001. Integrated multimedia processing for topic segmentation and classification. In *Proceedings of IEEE International Conference on Image Processing(ICIP2003)*, pages 366–369.

Daisuke Kawahara and Sadao Kurohashi. 2002. Fertilization of case frame dictionary for robust japanese case analysis. In *Proceedings of 19th COLING (COLING02)*, pages 425–431.

Daisuke Kawahara and Sadao Kurohashi. 2004. 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*, pages 334–341.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Huu Bach Nguyen, Koichi Shinoda, and Sadaoki Furui. 2005. Robust highlight extraction using multistream hidden markov models for baseball video. In *Proceedings of the International Conference on Image Processing 2005(ICIP2005)*, pages 173–176.

Dinh Q.Phung, Thi V.T Duong, Hung H.Bui, and S.Venkatesh. 2005. Topic transition detection using hierarchical hidden markov and semi-markov models. In *Proceedings of ACM International Conference on Multimedia(ACM-MM05)*, pages 6–11.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2004. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, number 1201–1207, 8.