Multimedia supplementation to a cooking recipe text for facilitating its understanding to inexperienced users

Ichiro Ide*¶†, Yuka Shidochi*‖, Yuichi Nakamura‡, Daisuke Deguchi*, Tomokazu Takahashi§ and Hiroshi Murase*

* *Graduate School of Information Science, Nagoya University*
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
E-mail: {ide@, shidochi@murase.m., ddeguchi@, murase@}is.nagoya-u.ac.jp
¶ *Informatics Institute, University of Amsterdam*
Science park 904, 1098 XH Amsterdam, The Netherlands
E-mail: I.Ide@uva.nl
† *National Institute of Informatics, Research Organization of Information and Systems*
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
‖ *Currently at Toyota Motor Corp., Japan*
‡ *Academic Center for Computing and Media Sciences, Kyoto University*
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
E-mail: yuichi@media.kyoto-u.ac.jp
§ *Department of Economics and Information, Gifu Shotoku Gakuen University*
1-38 Naka-Uzura, Gifu 500-8288, Japan
E-mail: ttakahashi@gifu.shotoku.ac.jp

***Abstract*—Assisting culinary activities for inexperienced users has been considered as an important task in most existing works in the field. On the other hand, recipe texts are becoming available on the Internet in increasing numbers. However, they tend to be written simply by mostly non-professional people, and thus are sometimes difficult for an inexperienced person to follow the steps and manage to cook as they are supposed to. In this paper, we propose a method that detects difficult descriptions for an inexperienced user in an existing text recipe, and supplements them with multimedia contents including text information extracted from a large number of recipes, and also images and video clips on certain kinds of cooking operations, to facilitate the understanding of the recipe. Experimental results showed promising ability of the proposed method to assist inexperienced users understand the descriptions in a recipe.**

***Keywords*-cooking recipe, description, rewriting, multimediatization**

I. Introduction

Supporting domestic activities inside a household by means of information and communication technology is becoming a realistic research topic. Among various possibilities, we are focusing on the culinary activity in a kitchen, which is a highly intellectual and creative activity that requires abundant experience and knowledge on the task. On the other hand, such a requirement poses difficulty for an inexperienced person to cook like an experienced person.

Considering this problem, several attempts have been made in the past decade to assist culinary activities in the kitchen, mostly targeting inexperienced users. Hamada et al. proposed the "Cooking Navi" system that analyzes the dependency structure in a cooking recipe text [1], align each step to a video segment obtained from corresponding cook shows [2], and also present the steps along the dependency structure with the aligned video while a user cooks [3]. Hashimoto et al. are working on the "Smart Kitchen" project [4] that assists cooking by detecting a user's activity from various sensors in a kitchen. On the other hand, a cooking assistance software for a portable game machine, "Nintendo DS" is already commercially available[1]. However, all of them simply facilitates the understanding of a text recipe by allowing a user to selectively access multimedia information on a fixed recipe.

Meanwhile, the number of cooking recipe texts posted on the Web is increasing. For example, "Cookpad"[2] is a recipe-based social networking service where users can post original recipes and report results and comments. It is so popular that it is said that one fourth of Japanese women in their thirties accesses this service. However, the recipes posted by general people tend to be relatively simple and not professionally edited, which sometimes makes it difficult for an inexperienced person to follow the steps and manage to cook as they are supposed to.

In this paper, in order to solve these problems, we propose a method that first detects difficult descriptions for an inexperienced user in an existing text recipe. It then supplements them with multimedia contents including text information

[1]Nintendo Co., Ltd., "It talks! DS Cooking Navi (in Japanese)," http://www.nintendo.co.jp/ds/a4vj/.

[2]COOKPAD Inc., "COOKPAD," http://cookpad.com/.

978-0-7695-4217-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ISM.2010.42

extracted from a large number of recipes, and also images and video clips on certain kinds of cooking operations.

Similarly, Miyawaki and Sano have proposed a cooking assistance system for users with higher brain dysfunction [5]. Their approach is similar to ours in the sense that it rewrites a recipe based on the user's needs. However, the approach decomposes a long description and expands an abbreviated description within a recipe, whereas our approach supplements a recipe by making use of external contents and general knowledge obtained from a large amount of recipes. We have also proposed a method that suggests the replacement of materials in a recipe according to a list of replaceable materials extracted from a large number of recipes [6], but it does not supplement descriptions as it does in the work presented in this paper. Some groups who participated in the "Computer Cooking Contest" series[3] have proposed methods that replace materials in a recipe together with the corresponding descriptions such as the cooking operations and quantities [7], [8], but these works do not focus on facilitating the descriptions according to users' knowledge and skills, nor supplementation by multimedia contents other than text.

Although the proposed method targets Japanese text, it should be able to be applied to other languages by a similar approach. Accordingly, some language-specific details are omitted in the following descriptions for the sake of simplicity.

II. Study on the Characteristics of a Recipe for Inexperienced Users

In order to clarify the points that should be considered in a recipe for inexperienced people, we analyzed 24 pairs of recipes provided from a professional recipe site[4]. Each pair is composed of a recipe for children and general users on a same dish, both described by text and illustrations. Here, we considered recipes for children as those carefully intended for inexperienced people, and recipes for general users as those for experienced people.

A. Characteristics of text description

The study revealed the following tendencies for text descriptions:

- Detailed description on the material itself and the operation applied to it, when handling a single material.
- Detailed description on the current state or the change of the state of the materials, when handling multiple materials.

[3]Held in conjunction with ECCBR2008, ICCBR2009, and 2010 conferences. The latest one was CCC2010 (http://vm.liris.cnrs.fr/ccc2010/).

[4]L-NET CO., LTD., "Bob & Angie," http://www.bob-an.com/.

Table I
Roles of image and video descriptions according to the types of cooking operations.

Cooking operation	Image	Sequential image	Video clip	Total
Mixing	6	0	18	24
Heating	17	0	12	29
Cutting	13	30	5	48
Decorating	3	0	0	3
Dipping	0	0	0	0
Cooling	2	1	4	7
Separating	2	0	0	2
Others	27	5	8	40
Total	70	36	47	153

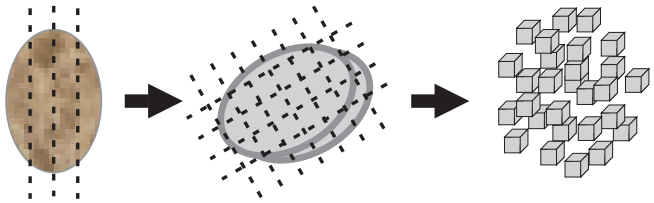

Figure 1. Example of a sequential image used to describe a cooking operation. In this case, "Cut a potato into small blocks."

B. Characteristics of image and video description

As shown in Table I, the following tendencies were observed. We followed the classification of cooking operations proposed by Hamada et al. [1] when analyzing the tendencies.

- Images supplement descriptions on "heating" operations.
- Sequential images supplement descriptions on "cutting" operations. As exemplified in Fig. 1, a "sequential image" is an image composed of multiple sub-images that depict different stages of a cooking operation.
- Video clips supplement descriptions on "mixing" operations.

Other operations did not show a strong tendency.

Based on these results, the proposed method first detects text descriptions that need to be supplemented, and then facilitates the descriptions by supplementing text or relevant image / video contents.

III. Supplementing the Difficult Descriptions for Facilitating Their Understanding

A. Detection of difficult descriptions

Based on the study in Section II, a pair of two terms that satisfy the following conditions are considered as difficult descriptions, and detected based on their general term frequencies.

1) Make a pair of "cooking operation (verb)" and either a "material (noun)," an "adjective," or an "adverb," based on the method proposed in our previous

work [6]. A Japanese morphological analyzer MeCab[5] was used to obtain the parts-of-speech of the terms.

2) Based on the general term frequency f_v and f_t for terms t_v and t_t that compose the pair, calculate the distinctiveness $D(t_v, t_t)$ of the pair as follows:

$$D(t_v, t_t) = \frac{1}{w_v f_v + w_t f_t} \quad (1)$$

Here, the weights w_v and w_t were empirically set as follows based on the study in Section II, according to the type of the "cooking operation" t_v.

- $(w_v, w_t) = (0.1, 0.9)$ when t_v is a verb that handles a single material.
- $(w_v, w_t) = (0.3, 0.7)$ when t_v is a verb that handles multiple materials.

3) Detect the pair (t_v, t_t) as a difficult description if $D(t_v, t_t)$ is larger than a threshold θ_D.

The terms were classified into "cooking operation (single)", "cooking operation (multiple)", "material" and counted based on dictionaries compiled beforehand by the following means:

- "Cooking operation (single, multiple)": Classified based on the dictionary compiled by Hamada et al. [1], and by counting corresponding verbs from the preparation steps in a large number of recipe texts.
- "Material": Collected and counted from the material list in a large number of recipes texts.

We collected and counted the frequency of 2,202 kinds of materials (nouns), 1,568 kinds of cooking operations (verbs; 795 kinds of single and 773 kinds of multiple), 197 kinds of adverbs, 118 kinds of adjectives, and 7,574 kinds of other nouns, from 6,779 recipes obtained from an online recipe site[6].

B. Preparation of supplementary contents

In order to supplement a difficult description, it is necessary to prepare the supplementary contents beforehand. Note that the classification of terms, and the dictionaries used are the same as those described in Section III-A.

1) Text-based supplementary contents: We considered that when a material is cooked in a similar process in many recipes, a common modifier that appears in a majority of them could be a general description for an operation to the material. Based on this idea, common descriptions t_{m_i} for a "material"–"cooking operation" pair (t_n, t_v) that appears in an input recipe text are extracted on the fly by the following steps:

1) Gather from a large number of recipes, recipes that contain a material t_n cooked in a similar process to the input recipe. Here, a cooking process is represented as a sequential list of "cooking operations (verbs)" $P(t_n, r_i) = \{t_{v_1}, t_{v_2}, ...\}$ in recipe r_i. The similarity of cooking processes for t_n in two recipes r_1 and r_2 is measured by a normalized edit distance of terms $d(P(t_n, r_1), P(t_n, r_2))$ defined as follows:

$$d(P(t_n, r_1), P(t_n, r_2)) = \frac{d_e(P(t_n, r_1), P(t_n, r_2))}{max(\|P(t_n, r_1)\|, \|P(t_n, r_2)\|)} \quad (2)$$

Where $d_e(P(t_n, r_1), P(t_n, r_2))$ represents the edit distance between the two sequential lists obtained by Dynamic Time Warping. When the distance is shorter than a threshold θ_d, the two recipes are considered similar from the point of the operations applied to t_n. The set of gathered recipes will be called a Common Cooking Operation Recipe Set (CCORS) for material t_n hereafter.

2) Extract modifiers t_{m_i} that appear in between all the (t_n, t_v) pairs in the CCORS. Here, a modifier is defined as all terms other than particles or auxiliary verbs. A modifier will typically be an adverb or an adjective.

3) For all t_{m_i}, measure the commonality $C(t_{m_i}; t_n, t_v)$ as follows:

$$C(t_{m_i}; t_n, t_v) = \frac{\text{Frequency of the cooccurence of } t_n, t_{m_i}, \text{ and } t_v}{\text{Frequency of the cooccurence of } t_n \text{ and } t_v} \quad (3)$$

4) Select modifiers with commonalities larger than θ_C as supplementary contents. Nonetheless, quantitative descriptions regarding quantity, time, and temperature are ignored, since they are highly sensitive to the context in each recipe.

2) Image / video-based supplementary contents: In addition to the text-based supplementary contents, image-based and video-based supplementary contents are also important to facilitate the understanding of cooking operations, especially for operations such as those introduced in Section II-B. In this paper, we consider that such contents are readily labeled and available from the Internet or broadcast TV archives based on methods such as those proposed in references [9] and [10]. In the following experiments, we collected such contents from the Internet and labeled them manually.

C. Multimedia supplementation of the descriptions

1) Text-based supplementation: A "material"–"cooking operation" pair (t_n, t_v) that appears in an input recipe text is supplemented with a description t_m if a triplet (t_n, t_m, t_v) is available in the supplementary contents obtained in Section III-B.

[5]Kyoto University, "Japanese morphological analyzer MeCab," http://mecab.sourceforge.net/.

[6]Ajinomoto Co., Inc., "Encyclopedia of Recipes (in Japanese)," http://www.ajinomoto.co.jp/recipe/.

Table II
EVALUATION OF THE DIFFICULT DESCRIPTIONS DETECTED BY EACH SUBJECT.

Subject	Recall		Precision	
1	55%	(18 / 33)	29%	(18 / 62)
2	76%	(13 / 17)	21%	(13 / 62)
3	60%	(13 / 22)	21%	(13 / 62)
4	74%	(14 / 19)	23%	(14 / 62)
5	85%	(11 / 13)	18%	(11 / 62)
6	82%	(9 / 11)	15%	(9 / 62)
7	61%	(25 / 41)	40%	(25 / 62)
8	60%	(24 / 40)	39%	(24 / 62)
9	88%	(7 / 8)	11%	(7 / 62)
10	64%	(18 / 28)	29%	(18 / 62)
11	64%	(9 / 14)	15%	(9 / 62)
12	63%	(25 / 40)	40%	(25 / 62)
Average	69%	—	25%	—

2) Image / video-based supplementation: Based on the study in Section II-B, an image, an image sequence, or a video clip corresponding to a "material"–"cooking operation" pair is supplemented according to the following rules. Note that in the recipes we studied, images were illustrations, but we consider that photographic images could also be used for the same purpose.

- An image supplements a description on "heating" operations.
- A sequential image supplements a description on "cutting" operations.
- A video clip supplements a description on "mixing" operations.

IV. EXPERIMENT

A. Detection of difficult descriptions

1) Setting: In order to evaluate the detection method of difficult descriptions, we compared the result of a subjective experiment and the output of the method proposed in Section III-A.

Eleven recipes obtained from an online recipe site[4] were used for the experiment. These were selected from recipes for experienced users even among the recipes for adults. The threshold θ_D was set so that the descriptions that were different between the recipes for general users and children in the analysis in Section II should be detected the most.

Twelve subjects who do not cook daily, and thus considered as inexperienced users, participated in the experiment. They were shown all the eleven recipes, and asked to mark all the descriptions that they considered difficult.

2) Result: The system detected 62 descriptions as difficult. Table II shows the evaluation of the result per subject. In average, 69% of the descriptions that the subjects pointed out as difficult to understand were correctly detected (recall), and 25% of the detected descriptions matched the users' (precision).

As we carefully analyzed the descriptions pointed-out by the subjects, they were quite different among the subjects. This indicates the necessity of personalization of the detection method in order to increase the recall for some users and also to increase the overall precision.

Table III
EXAMPLE OF THE EXTRACTED SUPPLEMENTARY DESCRIPTIONS.

Recipe	Material (t_n)	Cooking operation (t_v)	Supplementary description (t_m)
Stew	potato	peel	skin
	potato	cut	small blocks
	carrot	peel	skin
Croquette	flour	shake down	excessive
	egg	crack and mix	bowl
	cabbage	tear	hand
	onion	fry	pan
Cake	flour	mix	bowl
	butter	bake	oven

B. Extraction of supplementary descriptions

1) Setting: In order to evaluate the extracted supplementary descriptions, we manually analyzed the output of the method proposed in Section III-B.

Three recipes ("Japanese meat and potato stew," "Soybean croquette," and "Country apple cake") among the 6,779 recipes used to compile the dictionaries in Section III-A were used as input recipes. The thresholds were set to $\theta_d = 0.8$ and $\theta_C = 0.5$.

2) Result: In total, 185 supplementary descriptions were extracted. Among them, 136 descriptions (73.5%) were manually judged as appropriate. Table III shows an example of the appropriate supplementary descriptions.

As we analyzed the results, we found that the size of CCORS controlled by the threshold θ_d affected the result; if the size of a CCORS was large, the extracted supplementary descriptions were noisy, and if it was small, very few or even no description was extracted. This indicates the need to adaptively adjust θ_d according to each material.

C. Multimedia supplementation of actual recipe texts

1) Setting: Finally, we actually modified four recipes ("Japanese meat and potato stew," "Country apple cake," "Dorade steamed with vegetables" and "Beignet of shrimp and kidney pea") according to the method proposed in Section III-C, and had it evaluated in a subjective experiment. Figure 2 shows one of the modified recipes.

Some parts of the original recipes were modified to make it simpler, and then compared with the supplemented recipe. When the supplementary descriptions were inserted in the original recipe, surrounding texts were manually adjusted so that it becomes a natural sentence.

Eight subjects who do not cook daily, and thus considered as inexperienced users, participated in the experiment.

2) Result: Table IV shows the result of the experiment. Although most subjects judged that the supplemented recipe was better, there were some cases that it was about the same or even worse. In these cases, the subjects commented

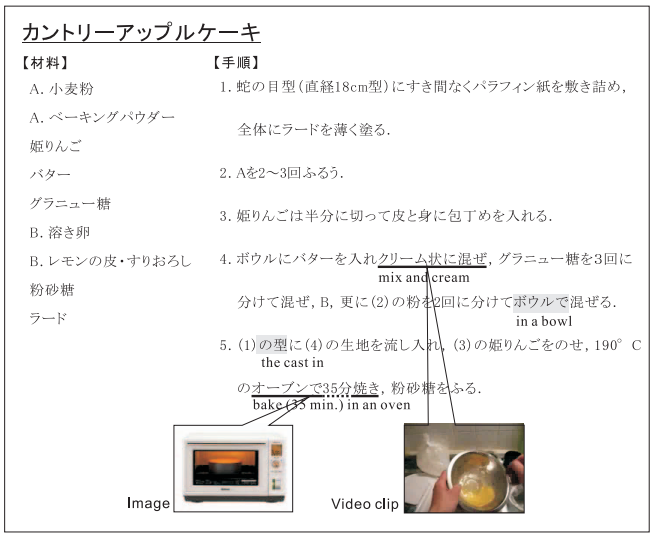

Figure 2. Example of an actual recipe supplemented by multimedia descriptions obtained from the proposed method. In this case, "Country apple cake." The highlighted text descriptions, the popuped image, and the video clip were supplemented. The text recipe was obtained from Ajinomoto's "Encyclopedia of recipes"[6].

Table IV
EVALUATION OF THE ACTUAL RECIPE TEXTS SUPPLEMENTED WITH MULTIMEDIA CONTENTS. FREQUENCY OF SUBJECTS WHO CONSIDERED EITHER OF THE RECIPES EASIER TO UNDERSTAND.

Recipe	Original is better	Supplemented is better	Neutral
Stew	0	8	0
Cake	0	6	2
Steamed dorade	0	8	0
Beignet	1	7	0

that some supplementations, especially when the name of a receptacle was supplemented, they made the description redundant, and even difficult to read.

Another major comment was that there were still descriptions difficult to understand. The difficult descriptions pointed out were different among subjects, so this again, confirms the necessity to adaptively handle personal difference as we noticed in Section IV-A.

V. CONCLUSION

In this paper, we proposed a method that detects and supplements difficult descriptions for an inexperienced user in an existing text recipe with text, image and video contents. Experimental results showed promising ability of the proposed method to assist users understand the descriptions in a recipe.

We are currently implementing a prototype interface that supplements existing recipes provided from a user [11]. We would also like to implement a scheme that allows us to adaptively handle the personal difference of experience and knowledge in the future. Detection and supplementation of difficult expressions even for experienced users (i.e. expressions that scarcely appear in most of the recipes) are also included in the future work.

ACKNOWLEDGMENT

Parts of this work were supported by the Grants-in-Aid for Scientific Research (21013022) from the Japanese Ministry of Education, Culture, Sports, Science and Technology. We would like to thank the subjects who participated in the experiments, and also Ms. Kanako Obata for joining the discussions.

REFERENCES

[1] R. Hamada, I. Ide, S. Sakai, and H. Tanaka, "Structural analysis of preparation steps on supplementary documents of cultural TV programs," in *Proceedings of the Fourth International Workshop on Information Retrieval with Asian Languages (IRAL'99)*, Taipei, Taiwan, Nov. 1999, pp. 43–47.

[2] R. Hamada, K. Miura, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, "Multimedia integration for cooking video indexing," in *Advances in Multimedia Information Processing —PCM2004 Fifth Pacific Rim Conference on Multimedia, Tokyo, Japan, November/December 2004, Proceedings Part II*, ser. Lecture Notes in Computer Science, K. Aizawa, Y. Nakamura, and S. Satoh, Eds. Berlin / Heidelberg, Germany: Springer-Verlag, Dec. 2004, vol. 3332, pp. 657–664.

[3] R. Hamada, J. Okabe, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, "Cooking Navi: Assistant for daily cooking in kitchen," in *Proceedings of the Thirteenth ACM International Multimedia Conference (ACMMM'05)*, Singapore, Nov. 2005, pp. 371–374.

[4] A. Hashimoto, N. Mori, T. Funatomi, Y. Yamakata, K. Kakusho, and M. Minoh, "Smart Kitchen: A user centric cooking support system," in *Proceedings of the Twelfth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'08)*, Málaga, Andalucía, Spain, Jun. 2008, pp. 848–854.

[5] K. Miyawaki and M. Sano, "A cooking support system for people with higher brain dysfunction," in *Proceedings of ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities (CEA'09)*, Beijing, China, Oct. 2009, pp. 47–52.

[6] Y. Shidochi, T. Takahashi, I. Ide, and H. Murase, "Finding replaceable materials in cooking recipe texts considering characteristic cooking actions," in *Proceedings of ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities (CEA'09)*, Beijing, China, Oct. 2009, pp. 9–14.

[7] A. Blansché, J. Cojan, V. Dufour-Lussier, J. Lieber, P. Molli, E. Nauer, H. Skaf-Molli, and Y. Toussaint, "Taaable 3: Adaptation of ingredient quantities and of textual preparations," in *Eighteenth International Conference on Case-Based Reasoning (ICCBR 2010) Workshop Proceedings*, Alessandria, Piemonte, Italy, Jul. 2010, pp. 189–198.

[8] M. Minor, R. Bergmann, S. Görg, and K. Walter, "Adaptation of cooking instructions following the workflow paradigm," in *Eighteenth International Conference on Case-Based Reasoning (ICCBR 2010) Workshop Proceedings*, Alessandria, Piemonte, Italy, Jul. 2010, pp. 199–208.

[9] R. Hamada, S. Satoh, and S. Sakai, "Detection of important segments in cooking videos," in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL) 2001*, Kauai, HI, USA, Dec. 2001, pp. 118–123.

[10] T. Shibata, N. Kato, and S. Kurohashi, "Automatic object model acquisition and object recognition by integrating linguistic and visual information," in *Proceedings of the Fifteenth International Conference on Multimedia (ACMMM'07)*, Augsburg, Bavaria, Germany, Sep. 2007, pp. 383–392.

[11] K. Doman, C.Y. Kuai, T. Takahashi, I. Ide, and H. Murase, "Video CooKing: Towards the synthesis of multimedia cooking recipes," in *Seventeenth International Conference on MultiMedia Modeling (MMM2011), Special Session on Multimedia Understanding for Consumer Electronics*, Taipei, Taiwan, to appear in Jan. 2011.