Complementation of Food and Tool Information in Multi-Modal Recipe Procedural Descriptions

Yixin ZHANG†, Yoko YAMAKATA††, and Keishi TAJIMA†

† Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan
E-mail: †zhangyx@dl.soc.i.kyoto-u.ac.jp, ††yamakata@mi.u-tokyo.ac.jp, †††tajima@i.kyoto-u.ac.jp

Abstract In this paper, we propose a method of complementing food and tool information omitted in procedural text descriptions of Chinese recipes when it is associated with illustration images. Tool and food omission sometimes makes the understanding of recipe text difficult. However, such omitted tool and food sometimes appear in the associated images. Therefore, our method automatically completes the omitted tool and food information by identifying tools and food in the images. Firstly, it recognizes tokens of food, tool, action performed by a chef and so on in a procedural text according to our previous research about recipe named entity. The bert-based named entity recognizer, "BERT-NER" achieved 91.3% of accuracy using 45 tagged recipes as training data. For word complement result, among all the 246,195 steps, there are 16,593 steps in which food information is omitted and 66,228 steps in which tool information is omitted. Our method achieves the accuracy of 83.26% for tool information complement and 43.57% for food.

Key words Recipe Data, Image Recognition, Recipe Named Entity, Word Complement

1 Introduction

In recent years, many user-submitted recipe sites, such as Allrecipes(注1) in North America and UK, Cookpad(注2) in Japan, and Haodou(注3) in China, have become popular. Nowadays several millions of recipe data posted by users are shared on such recipe sites. Each recipe data entry consists of the title of recipes, outlines, material and ingredient lists, cooking procedures, and tips. In some cases, images are associated with procedural steps in recipes. Figure 1 shows an example of such recipe data on Haodou Recipe Site. On this site each procedural step is associated with a corresponding image.

The data structure of recipe and their procedural steps are relatively uniform. In addition, this kind of recipes with one-to-one correspondence between images and procedural steps contains rich and valuable multimedia information. As a result, research on extracting and analyzing useful information in recipe data has become an important research issue in both natural language processing and computer vision field [1, 2].

Figure 1 Example of recipe data posted on Haodou

In this paper, we focus on the automatic complement of omitted tool and food information in text descriptions of

(注1)：https://www.allrecipes.com/
(注2)：http://cookpad.com/
(注3)：http://www.haodou.com/recipe/

recipe procedural steps by identifying tools and food in images associated with the corresponding procedural text.

When users upload recipes onto these websites, due to users' word habit, the text description sometimes has typos, infrequent expressions, and even omits some important information such as tool or food, which makes it difficult for machine such as smart speakers to understand. However, the omitted tool and food information in text description is sometimes shown in the associated images. We develop a method of complementing the text description of procedural steps by recognizing the images, and extracting features of tool and food from them.

The main processing flow is as follows.

First we detect sentences in which food or tool information is omitted by using Recipe Named Entity recognition [6]. We process text data as follows. A dataset of text-image pairs in procedural steps of recipes is constructed. We segment Chinese sentences into words and annotate words according to recipe named entities (r-NEs) [6]. Then we adopt the BERT-NER recognizer (NER) [7] to do the r-NEs recognition and train it using manually annotated 50 Chinese recipes. Our experimental results illustrate the high accuracy of our r-NEs recognition method of 91.27 %.

Second, for steps where food or tool information is omitted, we identify tool and food in the images in order to complement omitted words. Multi-label classification is a good method for parsing the information in the picture for many general tasks. We first assign labels of tool and food information to more than 500 images manually, and train a classifier on those images by using a standard multi-label classification method.

Then we complement the omitted food and tool information in the text of procedural steps using the label with the high probability score from the associated image.

Among the 12,548 recipes in our dataset, there are 5,685 distinct food names, where 24 food names appear more than 1,000 times, 298 food names more than 100 times but less than 1000 times, and 5,355 food names only 50 times or less, which accounts for 94.20% in total. There are also 1,620 distinct tool names. Among the 246,195 procedural steps, there are 16,593 steps in which food information is omitted and 66,228 steps in which tool information is omitted.

Therefore, we propose a method to complement food and tool information in procedural text descriptions. When food or tool information is omitted from the target step, the proposed method is used to recognize the corresponding image and assign the label with highest probability to the text.

2 Related Work

There have been many kinds of research on recipe text processing and image recognition. Currently, research on the recognition of images in cooking recipes mainly focuses on the whole dish appearance without explicit analysis of ingredient composition [2]. Ingredient and material estimation only from a completed food image is a task far harder than food categorization. Our method intends to recognize the food and tools omitted in the intermediate procedural steps, i.e., not in a final completed dish, in order to complement the needed information in text data.

There has also been research on the recipe text processing. Recipe text processing has some differences from general text processing, which makes it difficult to apply the existing text processing method easily to recipe data. [3]. A specific method for the recipe domain which even could be applied for the multilingual environment is desired.

The analysis of a set of words associated with images is also a research issue nowadays, such as tag identification from a tag set associated with an image [4] and inferring the semantic relationship between them [5]. They focus on the data from image posting sites like Flickr where the images are associated with tags already. In this paper, we use the text description along with associated images, and try to infer the relationship between images and text and do the complement work for the omitted fool and tools information.

In summary, recipe text processing and image recognition are important issues, and in this paper, we focus on the automatic complementing of food and tool information in procedural steps associated with images.

3 Proposed Method

We have explained briefly our methods of complementing omitted tools and food information in recipe data in the first section. In this section, we will explain some details of our proposed methods.

3 1 Dataset

We collected 12,548 recipe data posted on Haodou Recipe(注4), a user-submitted recipe site in China. Each data item consists of the following components: a recipe ID, a general description, ingredients, tips, and a sequence of cooking procedural steps, each of which is a pair of a text description and an associated image. Figure 1 shows an example.

We extract text data in procedural steps, segment Chinese sentences into words, and add Part-of-Speech tagging (POS tagging) by using Chinese language segmentation and POS tagging tool named jieba(注5). Then we constructed a Chinese recipe corpus of 50 recipes, and manually annotated

(注4)：http://www.haodou.com/recipe/
(注5)：https://github.com/fxsjy/jieba

Table 1 Recipe Named Entity (r-NE) Tags

Tag	Meaning	Remarks
Ac	Action by chef	Verb representing a chef ' s action
Ac2	Discontinuous Ac	Second, non-contiguous part of a single action by chef
F	Food	Eatable, also intermediate products
T	Tool	Knife, container, etc.
Sf	Food state	Food ' s initial or intermediate state
St	Tool state	Tool ' s initial or intermediate state
D	Duration	Duration of cooking
Q	Quantity	Quantity of food
At	Action by tool	Verb representing a tool ' s action
Af	Action by chef	Verb representing action of a food

recipe named entities (r-NEs) according to guidelines previously defined for Japanese and English [6]. Table 1 shows 10 types of r-NE and their meanings.

For image data, we process images by giving multi-label of tool and food information to images manually first, and use a standard multi-label classification method to recognize the probability of labels in the rest of the image data.

3 2 Recipe Named Entity in Chinese

First we are going to detect sentences in which food or tool information is omitted. We use the definition of Recipe Named Entity (r-NE), which is an example of a domain-specific NE definition for the recipe, to annotate the texts. Mori et al. constructed a Japanese recipe corpus consisting of 208 recipes randomly sampled from the Cookpad website [3] and Yamakata et al. extended and adopted it into English and added two more tags, Ac2 and At, in order to account for additional phenomena in English languages [6]. In this paper, we adopt this method to recipes procedural text in the Chinese language.

We first ask Chinese native speakers to annotate 50 recipes according to guidelines for English recipes. Then we train the named entity recognizer model named BERT-NER(注6) , a state-of-the-art named entity recognizer which is constructed based on the BERT neural network architecture [7] on the Chinese r-NEs for recipes in our dataset.

3 3 Recipe Image Recognition

We do the multi-label image classification by using Inception Net v3 [9], which is a deep convolutional neural network trained for single-label image classification and trained on ImageNet data(注7). First of all, we prepare the network with correct labels for each image in the training set. Then we retrain the last layer of the model and modify the method of evaluating generated predictions to be actually able to train it with regard to multiple possible correct classes for each image.

For the fully-connected layer, we replace the softmax function with sigmoid, in order to apply for the multi-label classification rather than the single-label case. We change the last layer, in which the resulting class probabilities could be able to express that an image belongs to different classes with different probabilities.

For the multi-labeling, based on the same image dataset, we have two methods to label the images: labeling food and tool information by a single model and labeling them separately. We compare two methods in the experiment and choose the one with better accuracy for further experiment.

3 4 Word Complement

We first traverse all the text data that has been tagged by BERT-NER. For such a procedural step which food or tool information is omitted, we assign the label with the highest probability score in the multi-label classification result in the associated image to these blanks, so that each step could have three basic attributes: Action (Ac), Food (F) and Tool (T). In this way, the recipe procedural step data with tool and food information complemented is provided.

4 Experiment

In this section, we explain the contents of the experiment and then discuss the results of the experiment in order to evaluate the proposed method.

4 1 Chinese r-NE Recognition Accuracy

To evaluate accuracy of the r-NE recognizer in Chinese, we process 2,142 tokens with 1,472 phrases, and the number of correct ones is 1,283. The results shows that the overall accuracy is 91.27% and for each entity type, the precision, recall and FB1 score are as shown in table 2. In this table, the entity types Af (Action by food) and At (Action by tool) were excluded because these entity types did not appear in the dataset.

Figure 2 shows the proportions of each r-NE tag in the Chinese corpus. From Figure 2 we can find that, besides the Ac(action verbs) tag, which we have analyzed in the previ-

Table 2 Recipe Named Entity (r-NE) Recognition Accuracy

Tag	Precision	Recall	FB1
Total	87.28%	87.16%	87.22
Ac	93.70%	92.94%	93.32
Ac2	68.97%	76.92%	72.73
F	83.15%	93.77%	90.34
T	87.10%	76.06%	81.20
Sf	73.23%	76.23%	74.70
St	58.82%	47.62%	52.63
D	92.31%	94.74%	93.51
Q	88.89%	100.00%	94.12

(注6) : https://github.com/kyzhouhzau/BERT-NER

(注7) : https://github.com/IntelAI/models/tree/master/benchmarks/imagerecognition/tensorflow/inceptionv3

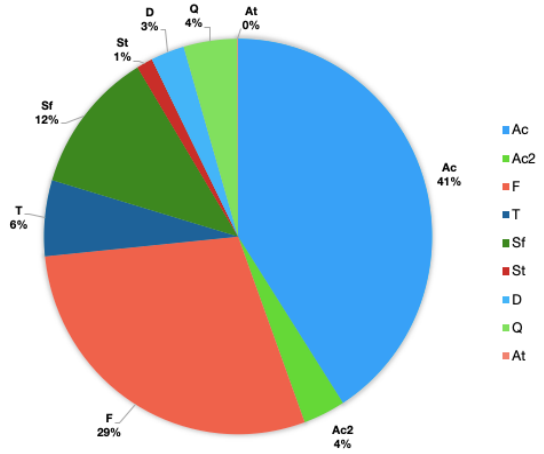


Figure 2 Proportions of the 10 r-NE tag types

ous research [10], food and tool tags have high proportions. Therefore, we could easily find that food and tool information may play an important role in the recipe data analysis, and if this kind of information is omitted, it will lead to great challenges in understanding the semantics for machines.

4 2 Image Recognition and Word Complement

We experiment with the two multi-label classification methods explained in 3 3 as shown in Figure 3. Method 1 is labeling food and tool information by a single model. In other words, food and tool are in the same label set. The second one is labeling them separately. We divide labels into two sets: tool label sets and food label sets.

We randomly select 500 pieces of data for traing, 150 pieces for validation and 500 pieces for testing. We first label these images manually and retrain the model using these data in two ways (single model and two separate models). As shown in Table 3, the accuracy of final test is over than 86% in both the two models.

Then among all the steps where food or tool information is omitted, we randomly select 3,000 pieces of procedural steps from both tool omitted and food omitted steps and adopt these two models to these data. The automatic-complemented result produced by our two methods are then compared with the result of manual annotation of omitted food and tool information in order to calculate the accuracy. As shown in Figure 4, Method 1, which is labeling them by a single model, achieves the accuracy of 82.13% for tool information complement and 42.63% for food. Method 2, which is labeling them by two separate models, among the steps selected in which the tool name is omitted, 83.26% of them are complemented correctly. However, among the steps selected in which the food name is omitted, 43.57% of them are complemented correctly.

In both cases, The accuracy for food is much lower than for tool. It is because when the multi-label classification is used to recognize the food information in images, the number of food labels is ultimately limited, and it is difficult to recognize the food names with low occurrences which make up the vast majority.

Food names and tool names in recipe text descriptions have different characteristics. For example, the number of the kinds of food is much bigger than that of tool. It may lead to the difference in difficulty of image feature recognition in these two cases. Therefore although both the accuracy of these two methods are similar, we choose to use the second method in order to complement tool and food information more efficiently, that is, use separate multi-label classification models for tool and food respectively. When tool information is omitted in the text of procedural steps, the multi-label classification model of the tool information is adopted. Similarly, when food information is omitted, the multi-label classification model of the food information is adopted.

By using the result of image recognition brought from the multi-label classification model, we could complement the omitted tool and food information.

From the experiment result, we can know that, among the 246,195 procedural steps in our dataset, there are 16,593 steps in which food information is omitted in text and 66,228 steps in which tool information is omitted in text.

We first assign labels of tool and food information to more than 500 images manually, and recognize the probability of labels in the rest of the image data using a standard multi-label classification method. Then we complement the omitted food and tools information in the text of procedural steps with the high probability label of associated images.

For example, in Figure 4(注8), the corresponding text of this image is "豆芽/**F** 摘洗/**Ac** 干 净/**Sf** ", which means "Pick/**Ac** and wash/**Ac** the sprouts/**F** clean/**Sf**", obviously the tool information is omitted, but in the image we could find out what the tool is. Through the recognition of the multi-label classification model, we can find that the label of the tool information with the highest probability is "plate", then we can add this information into this sentence. Similarly, the label with the second highest probability, "board" is also shown in the image.

On the other hand, in Figure 5(注9), the corresponding text of this image is "切片/**Ac** 腌制/**Ac**", which means "Cut into slices/**Ac** and Pickle/**Ac**", both the tool and food information are omitted. We need to adopt both the two models and take the labels with highest probability in these two models, that is, "small dish" and "meat". In this way, we can complete the original sentence with the tool information

(注8) : http://www.haodou.com/recipe/7003890
(注9) : http://www.haodou.com/recipe/7006984

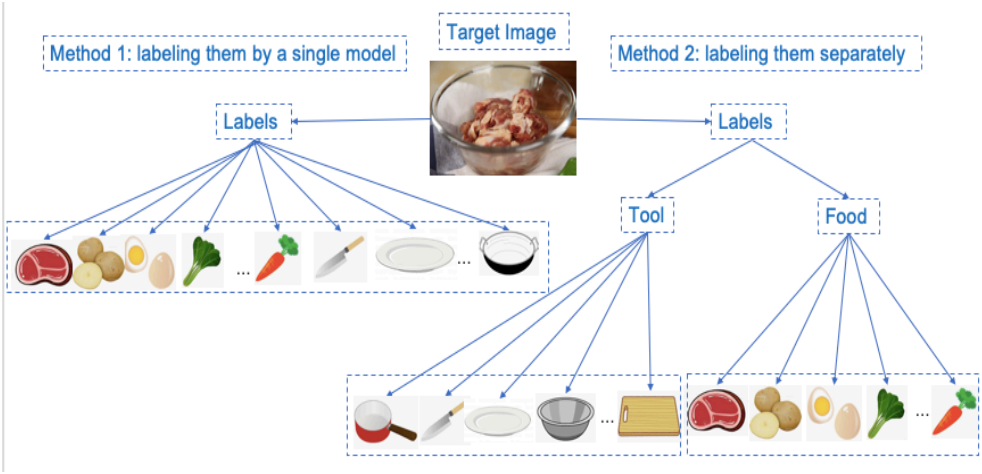


Figure 3 Two multi-label classification methods

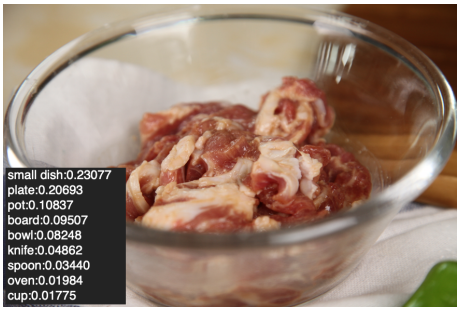
Table 3 Comparision of the two methods

	Labeling by a single model	Labeling them separately	
	Single model	Model of tool	Model of food
Final test accuracy	92.8%	86.4%	96.3%

Table 4 Compare the result of the two methods with the manual complement

	Labeling by a single model		Labeling them separately	
	Tool information	Food information	Tool information	Food information
Accuracy	82.13%	42.63%	83.26%	43.57%

Figure 4 Example of tool recognition result



(a) Tool recognition result

(b) Food recognition result

Figure 5 Example of food and tool recognition

"small dish/**T**" and the food information "meat/**F**".

Table 5 shows some positive examples of comparison before and after tool and food information complement.

On the other hand, there are also some negative examples in the experimental results in some cases. In the case of tool information, if tool does not exist in the image (Figure 6 (a)(注10)), or if the food is tool-shaped like plate (Figure 6 (b)(注11)), then misidentification may occur. In addition, tools with similar shape but different sizes sometimes also leads to confusion. The shape of the plate is similar to small dish, and sometimes it will be confused when the images with plate or small dish are recognized (Figure 6 (c)(注12)). In the case of food information, due to the variety of foods and their forms, misidentification sometimes occurs.

According to the accuracy and examples in practice, compared with food information, this method is better to adopt for recognizing the tool information in images. It also has some limitations. As mentioned above, there are some fac-

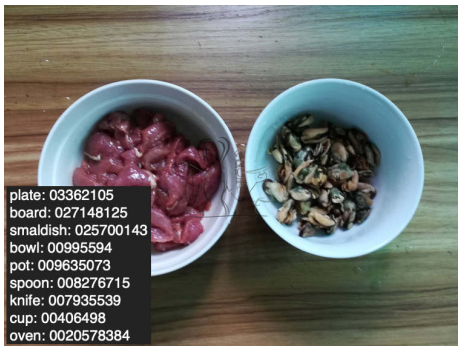
(注10)：http://www.haodou.com/recipe/7003897
(注11)：http://www.haodou.com/recipe/7003888
(注12)：http://www.haodou.com/recipe/7003890

Table 5 Example: Comparison of the r-NE before and after word complement

Example		F	AC	T
1	Before	鱼 Fish	腌制 Pickle	
	After	鱼 Fish	腌制 Pickle	碗 Bowl
2	Before	胡萝卜 Carrot	切开 Cut	
	After	胡萝卜 Carrot	切开 Cut	刀 Knife
3	Before		静置 Set Aside	碗 Bowl
	After	蛋液 Egg liquid	静置 Set Aside	碗 Bowl
4	Before		放入 Put in	烤箱 Oven
	After	面 团 Dough	放入 Put in	烤箱 Oven
5	Before		切碎 Finely chop	
	After	肉 Meat	切碎 Finely chop	案板 Board
6	Before		切碎 Finely chop	
	After	蛋 Egg	蒸熟 Steam	锅 Pot

(a)

(b)



(c)

Figure 6 Negative Examples of recognition

tors that can cause misidentification.

5 Conclusion

Action by chef, Food and Tool are three main entities in recipe procedural text, thus the complement of this kind of key information could help understand and enrich the content of recipe. In this paper, we propose a method of complementing food and tool information in procedural text descriptions.

We first adopt the method of Recipe Named Entity into Chinese recipe and construct the Chinese r-NE dataset and achieve 91.27% overall accuracy. Then the multi-label classification for image data and recognize the features in images is used to complement the food and tool information omitted from the procedural text description.

In general, the method proposed in this paper complements the omitted tool and food information to a certain extent. However, there are some factors that can cause misidentification. We will increase the accuracy of the experiment by increasing the size of training sets and types of labels, and along with method about contextual semantic relationships. The presence of multiple tools or foods also needs to be considered. In addition, in the future we will consider the situation where there are multiple tools or multiple foods in the picture. In addition, as we could find from the result, this method could be used to recognize and complement the tool information well but still has some limitations in food information recognition and complement.

In future work, besides the complement of tool information, we need to figure out a proper method for food information complement as well in order to improve the accuracy of complement. What is more, since a single procedural step associated with a single image often includes more than two action verbs. In such cases, we intend to determine which action verb the image is representing and complement the action verbs which do not be presented by the associated image. We would like to combine the prior work [10], which is about the automatic understanding of actions from recipe data consisting of procedural text and images, with the work in this paper, in order to generate recipes with more detailed and rich information in both text and image.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR16E3, JSPS KAKENHI Grant Number 18H03245, and JSPS KAKENHI Grant Number 18K11425, Japan.

References

[1] Wang, Xin, et al. "Recipe recognition with large multimodal food dataset." 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2015.

[2] Chen, Jingjing, and Chong-Wah Ngo. "Deep-based ingredient recognition for cooking recipe retrieval." Proceedings of the 24th ACM international conference on Multimedia. ACM, 2016.

[3] Mori, Shinsuke, et al. "Flow Graph Corpus from Recipe Texts." LREC. 2014.

[4] Li, Shangwen, et al. "Measuring and predicting tag importance for image retrieval." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2423-2436.

[5] Katsurai, Marie, Takahiro Ogawa, and Miki Haseyama. "A cross-modal approach for extracting semantic relationships between concepts using tagged images." IEEE Transactions on Multimedia 16.4 (2014): 1059-1074.

[6] Yamakata, Yoko, John Carroll, and Shinsuke Mori. "A comparison of cooking recipe named entities between Japanese and English." Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence. ACM, 2017.

[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[8] Sasada, Tetsuro, et al. "Named entity recognizer trainable from partially annotated data." Conference of the Pacific Association for Computational Linguistics. Springer, Singapore, 2015.

[9] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Zhang, Yixin, Yoko Yamakata, and Keishi Tajima. "Categorization of Cooking Actions Based on Textual/Visual Similarity." Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management. 2019.