Non-ingredient Detection in User-generated Recipes using the Sequence Tagging Approach

Yasuhiro Yamaguchi, Shintaro Inuzuka, Makoto Hiramatsu, and Jun Harashima
{yasuhiro-yamaguchi, shintaro-inuzuka, makoto-hiramatsu, jun-harashima}@cookpad.com
Cookpad Inc.
Yebisu Garden Place Tower 12F, 4-20-3 Ebisu, Shibuya-ku, Tokyo, 150-6012, Japan

Introduction

- At the present time, many people upload their recipes to the Internet, and over 6.7 million recipes have been uploaded to Cookpad, one of the largest recipe sharing services in the world
- However, some items in an ingredient list in a user-generated recipe are not actually edible ingredients in a user-generated recipe
- Such noise makes it difficult for computers to use recipes for a variety of tasks, such as calorie estimation
- We propose a method to detect non-ingredient items from an ingredient list in a recipe

Title	
ナスとピーマンの味噌炒め (Eggplant and Green Pepper Miso Stir-fry)	
Ingredient list	
ナス (eggplant)	5 個 (5 pieces)
ピーマン (green pepper)	5 個 (5 pieces)
調味料 (seasoning)	N/A
味噌 (miso)	大さじ 3 (3 tbs)
砂糖 (sugar)	大さじ 2 (2 tbs)
酒 (sake)	大さじ 2 (2 tbs)
Steps	
1. ナスを輪切りにする (cut eggplants into round slices)	
2. ...	

Figure 1:Example of a recipe. The N/A means that the user (i.e., recipe author) has not written the information. 調味料 (seasoning) is not an ingredient but the heading for the following ingredients.

Task Definition

The primary task in this study is to classify an item in an ingredient list as an ingredient or non-ingredient

Non-ingredient

- We define non-ingredient items based on edibility
- ``調味料'' (seasoning) is non-ingredient because it is used as a heading
- ``(↑ バターでもいいです)'' ((↑ you can use butter)) in Figure 2(a) is used as a comment, which mentions the previous ingredient ``マーガリン'' (margarine)
- ``竹串'' (bamboo skewers) in Figure 2(b) is a non-ingredient because it is not edible

Ingredient list	
...	...
マーガリン (margarine)	60g (60 grams)
(↑ バターでもいいです) ((↑ you can use butter))	N/A
砂糖 (sugar)	40g (40 grams)
...	...

(a) Example of a cookie recipe

Ingredient list	
...	...
サラダ油 (vegetable oil)	小さじ 1 (1 tsp)
竹串 (bamboo skewers)	3 本 (3 pieces)
岩塩 (rock salt)	適量 (desired amount)
...	...

(b) Example of a meat roll recipe

Figure 2:Examples of ingredient lists

Proposed Method

Ingredient Representation

- **TF-IDF**: We compute TF-IDF vectors for each item in the ingredient list

$$\mathrm{tf}(i,j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$\mathrm{idf}(j,D) = \log \frac{|D|}{|\{d \in D : t_i \in d\}|}, \quad (2)$$

$n_{i,j}$ is the number of words t_i in the j th ingredient name, d is the set of tokenized words in the ingredient name, and D is the set of all ingredient names in the recipe dataset

- **char-CNN**: Instead of TF-IDF, we can also use a CNN-based sequence encoder to obtain the character-level features of ingredient names

Model:

- We use the sequence tagging model shown in Figure 3
- By performing a non-ingredient detection task as a sequence tagging problem, the model can make predictions by taking items before and after the target item into account

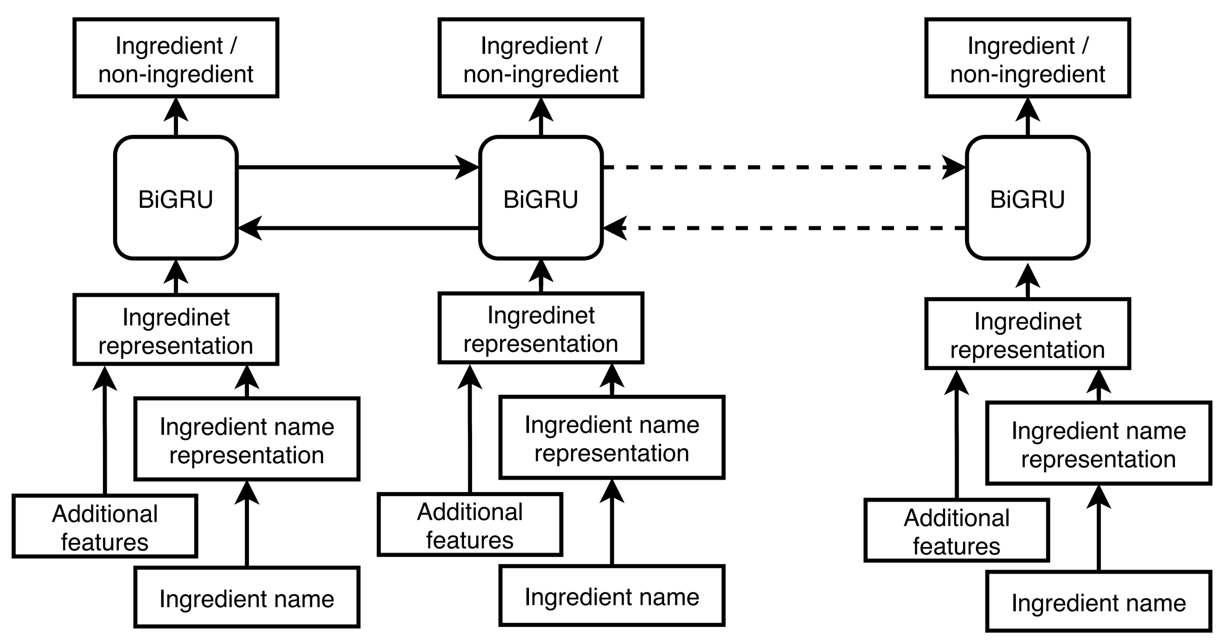

Figure 3:Overview of our method

Experiment

Dataset

- We chose 600 recipes from Cookpad
- # of items in these ingredient lists is 6675 (Ingredients: 5829 / Non-ingredients: 846)
- Each ingredient in the recipes was labeled as an ingredient or non-ingredient by three domain-expert annotators
- We collected recipes whose ingredient lists contained items without quantity information because such items tended to be non-ingredients in our preliminary investigation

Methods

- **Random forest** (baseline model): The input of the random forest model was the ingredient representation described in the previous section
- **BiGRU** (our model): We used two-layer bidirectional GRU (BiGRU) whose dimension of the hidden layer was 128

Results and Discussion

Results

- The BiGRU-based model was better than the random forest
- The sequence labeling approach was effective for the non-ingredient detection task

Model	F1	Precision	Recall
Random Forest	87.2 ± 5.1	82.8 ± 8.3	92.8 ± 4.5
+ ingredient freq.	88.8 ± 4.1	86.8 ± 6.6	91.4 ± 4.8
BiGRU + TF-IDF	90.8 ± 3.1	90.1 ± 3.7	91.2 ± 3.6
+ ingredient freq.	91.6 ± 3.4	89.4 ± 4.8	94.1 ± 3.6
BiGRU + char-CNN	91.2 ± 2.7	91.3 ± 4.8	91.4 ± 3.5
+ ingredient freq.	$\mathbf{93.3 \pm 2.3}$	$\mathbf{93.2 \pm 3.7}$	$\mathbf{94.1 \pm 3.1}$

Table 1:Experimental results

Discussion

- The ingredient frequency improved the F1 scores for both the random forest and BiGRU models
- As shown in Table 2, ingredient names occurred frequently in recipes were actual ingredients, so ingredient name frequency is important for ingredient detection
- The ingredient frequency can be an alternative feature of an ingredient dictionary which is usually rarely available

Name	Frequency
砂糖 (sugar)	524, 647
塩 (salt)	507, 766
水 (water)	450, 370
卵 (egg)	400, 572
醤油 (soy sauce)	320, 834

Table 2:Top 5 ingredient names in our dataset

Conclusion

- We introduced a non-ingredient detection task for user-generated recipes and proposed a neural model based on the sequence tagging approach
- We used a BiGRU-based model to predict a label for each ingredient over an ingredient sequence
- To evaluate our method, we constructed a dataset that contained 6,675 ingredients of 600 recipes from Cookpad
- Our experimental results showed that the proposed method achieved a 93.3 F1 score in the task
- In future work, we plan to verify the effectiveness of our method for downstream tasks, such as calorie estimation