Supplementing Omitted Named Entities in Cooking Procedural Text with Attached Images

Yixin Zhang¹, Yoko Yamakata² and Keishi Tajima¹
¹Kyoto University ²The University of Tokyo

MIPR 2021
September 9, 2021
Research Background

• In recent years, user-submitted recipe sites have become popular.

• Recipes with one-to-one correspondence between images and texts

• Recipe website: Haodou

source: https://www.haodou.com/recipe/1190778
Research Problem

Example:

• when people interacting with smart speakers...

  ![Image of a vegetable](image)
  
  Cut into slices.
  
  Food entity is omitted in the text.

• **Foodstuffs are omitted** in some instructional text.
  
  • This can make the recipe or a particular procedure difficult to understand.
  
  • However, those entities omitted in text are sometimes shown in the **instructional images**.

• If we want to **supplement food in text**, we need to **recognize food** in instructional images.
Research Problem

Depending on the situation, the **appearance** of the **same ingredients** (in the images) may **differ**.

On the other hand, even if the **images are similar**, the **food (in the text) might be different** in different situations.

- Prepare two potatoes.

- Cut them into pieces.

- Boil for 5 minutes.

Beginning Phase

Intermediate Phase

Finishing Phase

Potato

Spring onion

Beans
To supplement food in text,

we propose a method ——

recognizing food entity candidates based on both sentence similarity and image similarity
Food Recognition Based on Similarities

Target text-image pair

- Cut into slices.

Food is omitted in the sentence (Want to supplement)

Text-image pair dataset

- Chop the onion.
- Prepare a onion.
- Cut the meat.

Compute the similarity

- Text similarity
- Image similarity

Obtain the candidates with highest probability

Chop the onion.
Examples of Experiment Results

- The example results of the sentence embedding method for calculating text similarity

Target text-image pair

Quickly add egg pieces.
Add egg pieces.
Pour in egg pieces and stir fry.
Pour in egg pieces.
Add egg pieces and stir fry.
Add egg pieces and fry for a while.
Wash the eggs.
Put in egg pieces

Text-image pair set with high similarity to the target.

Stir fry well and stand by
Evaluation

• Evaluation: compare the result with manually labeled results and compute the intersection.

• Our method: 67.55%
  • Can supplement arbitrary food classes appearing in the dataset
• (Baseline) Ordinary Inception V3: 43.57%
  • Cannot supplement food that are not included in the 10 classes over which the model is trained.
Thanks for listening!