Supplementing Omitted Named Entities in Cooking Procedural Text with Attached Images

Yixin Zhang[1], Yoko Yamakata[2] and Keishi Tajima[1]
[1]Kyoto University and [2]The University of Tokyo, Japan

RESEARCH BACKGROUND

- In recent years, user-submitted recipe sites have become popular.
- Recipes with text-image paired instructions.
- Recipe website: Haodou

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

RESEARCH PROBLEM

Example: when people interacting with smart speakers...

Food entity is omitted in the text.

- **Foodstuffs are omitted** in some instructional steps.
  - difficult to understand.
- However, those omitted entities in text descriptions are sometimes shown in the **attached images**.

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

- The appearance of the same ingredients (in the images) may differ.

potato

Prepare two potatoes. | Cut them into pieces. | Boil for 5 minutes.

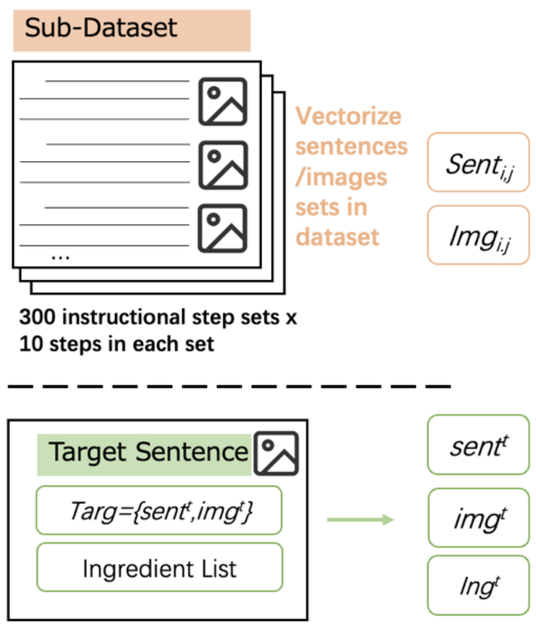
- Even if the images are similar, the food (in the text) might be different in different situations.

Potato | Spring onion | Beans

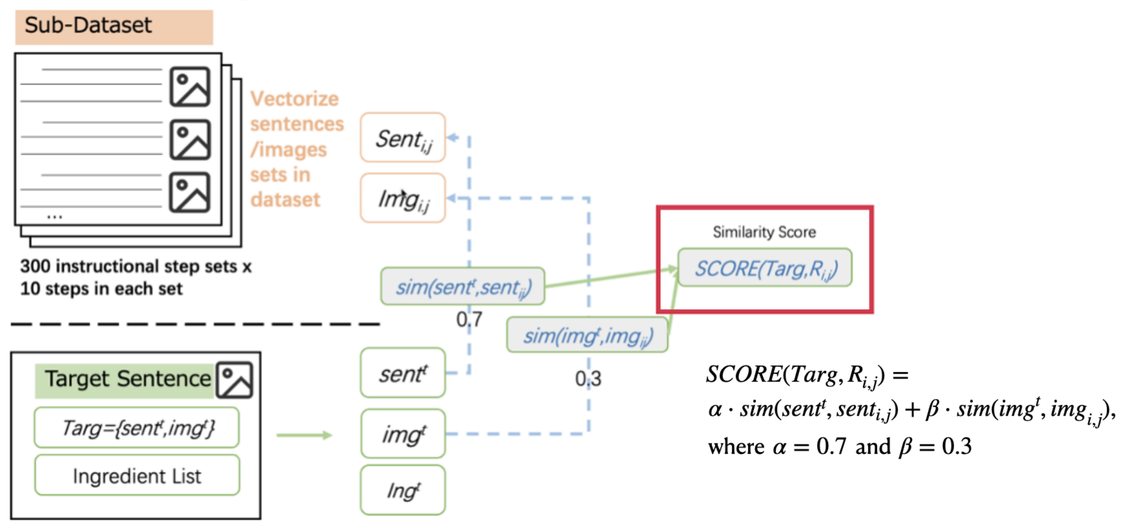
PROPOSED METHOD - FOOD RECOGNITION BASED ON SIMILARITIES

- Food Recognition Based on **Sentence Similarity and Image Similarity**
- Text data:
- Sentence2vec → 100-D vector
- Image data: VGG16fc-layer → 4096-D vector.

- Sub-database:
  - Select 300 representative instructional sentences with images from the recipe dataset.
  - For each sentence:
    - select 10 sentences that are similar to it
    - obtain the ten pairs of sentence and image.



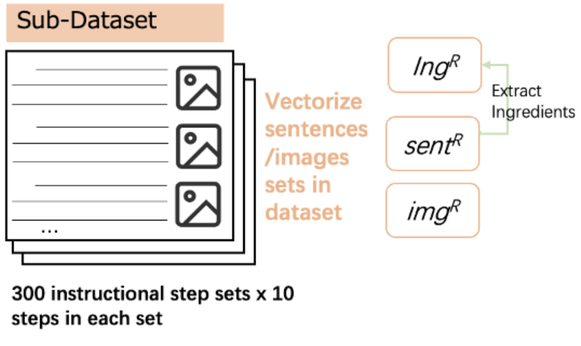
- From the sub-dataset, select the sentence/image pair that is closest to the target.
- The score is calculated by integrating the similarity of both the sentence and image vectors



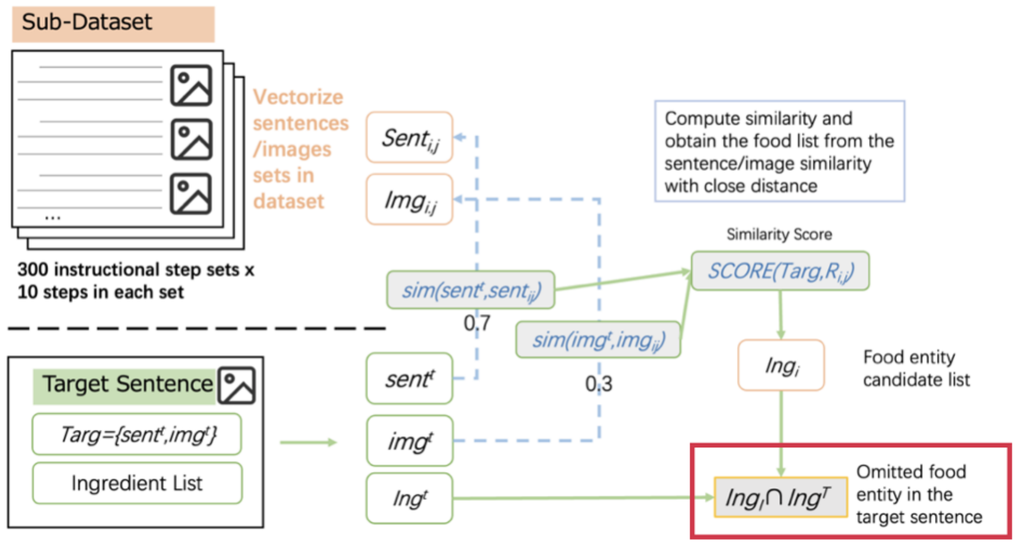
- From each representative sentence and its similar ten sentences, a set of ingredients contained in those sentences is extracted.
  - The purpose: collect the set of ingredients

↓

can be a target of the cooking steps expressed in similar sentences.



- At the last, the omitted food entity in the target sentence is supplemented with the ingredients from this intersect.



EXPERIMENT AND EVALUATION

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

Target text-image pair

Stir fry well and stand by

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

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

- 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.

CONCLUSION

- **Contributions:**
1. We construct a dataset of Chinese recipes consisting of 12,548 recipes.
2. We develop a recipe Named Entity (r-NE) recognizer [14] in Chinese
3. To solve the difficulty of recognizing food in different cooking stages, we propose a method of obtaining food entity candidates from other steps that are similar to the target step, both in sentence similarity and image feature similarity.

- Future work
  - Since food states change over time during cooking, we could use this feature to greater effect in order to improve the identification of food.
  - We would like to use the relationship between text and images to enrich the information content and structure of recipes so as to be more conducive to the application of recipe retrieval or automatic translation.