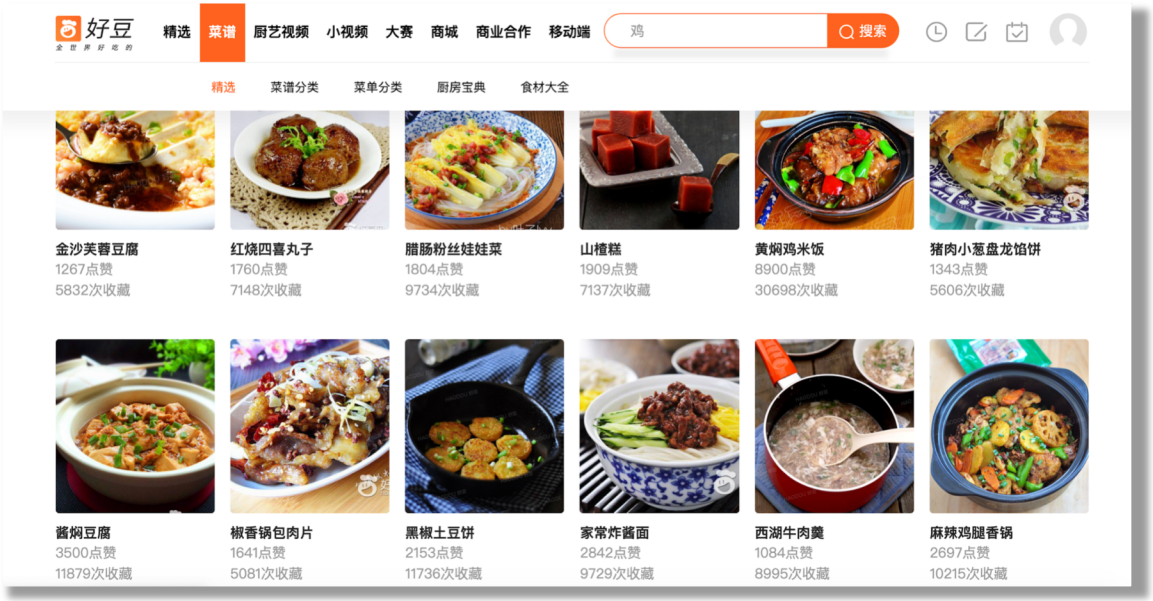
Categorization of Cooking Actions Based on Textual/Visual Similarity

Yixin Zhang[1], Yoko Yamakata[2] and Keishi Tajima[1]

[1]Kyoto University and [2]The University of Tokyo, Japan

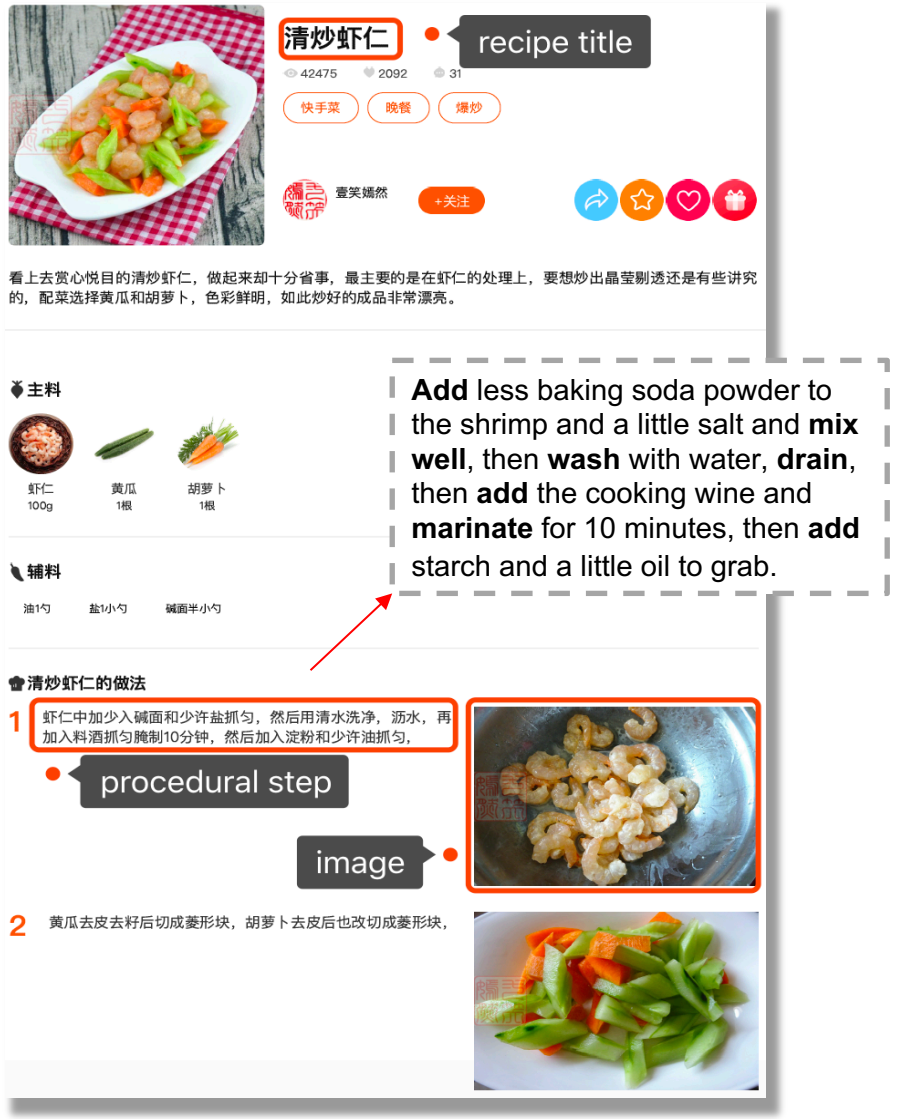
MADiMa 2019

Research Background



- 5,482,309 recipes in Chinese
- Every procedural step is associated with an image.
- The pairs of text and image in procedural steps are used in this research.
- 57,361,678 steps in total

Chinese recipe site: 好豆 (Haodou)



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

Research Purpose

Word embedding of action verbs in recipes.

- To calculate similarities and differences between action verbs.
  - “Cut” and “Cut into” have similar meanings but the latter one indicates the shape of ingredients after being cut.
- To be used for recipe retrieval or recipe automatic translation.

- Existing method: word2vec
  - Train word-embedding model by recipe text data and transform each verb into a vector.
  - The similarity between two verbs is computed by the cosine similarity of their embedded vectors.

Basic observation

1 青椒洗净后切丝

cut

Wash the green pepper and **cut**.

1 将茄子切成粗条。

cut into

Cut eggplant **into** sticks.

5 骨汤倒入锅中，大火翻炒均匀，调入盐淋入核桃油即可。

stir fry

Pour the bone soup into the pot, **stir fry** over high heat, and mix in salt and pour in the walnut oil.

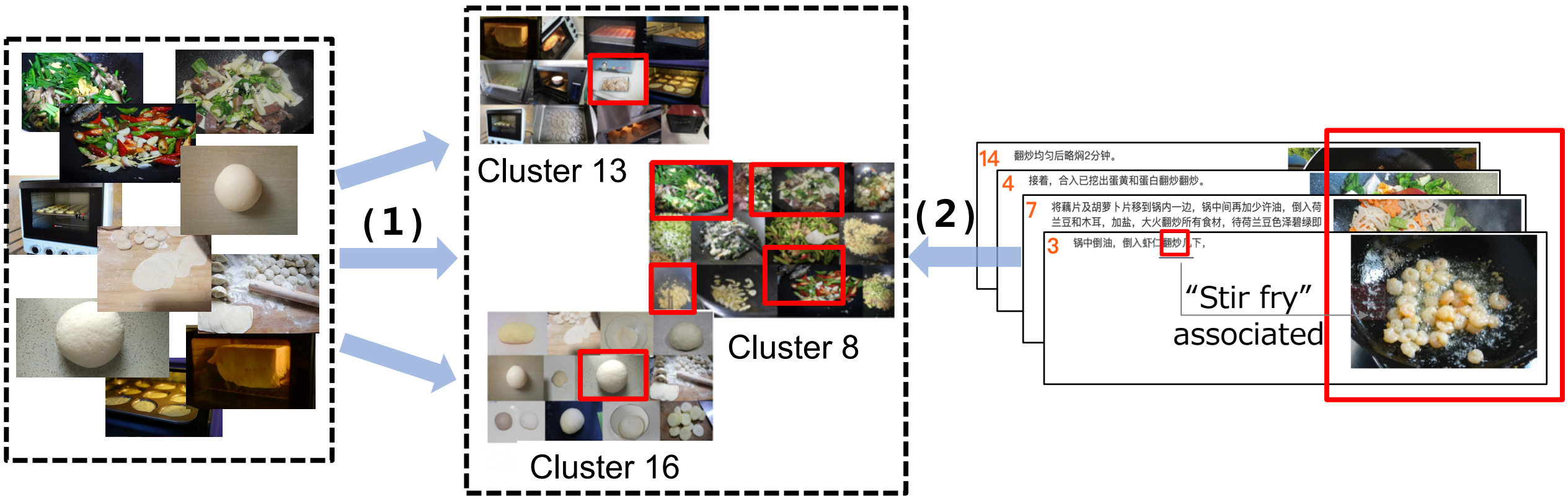
The semantic of “cut” and “cut into” are similar, also the image styles of them are same (ingredients on the chopping board)

On the other hand, the semantic of “cut” and “stir fry” are not similar, also the image styles of them are not similar (latter one: ingredients in the fry pan)

Proposed Method

- Word embedding by associated images.

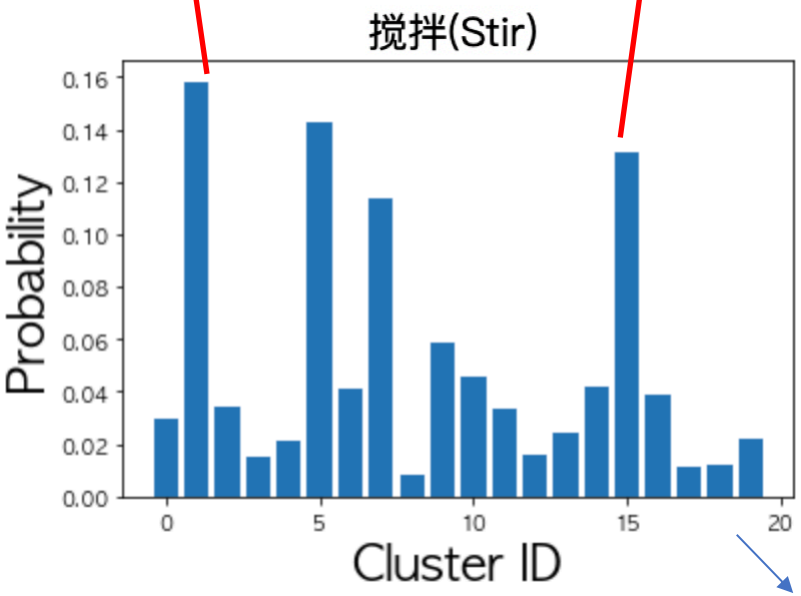
  (1) Clustering images in our dataset into groups which consist of images of similar style.
  (2) For each verb, calculating the probability distribution of the associated images over clusters.

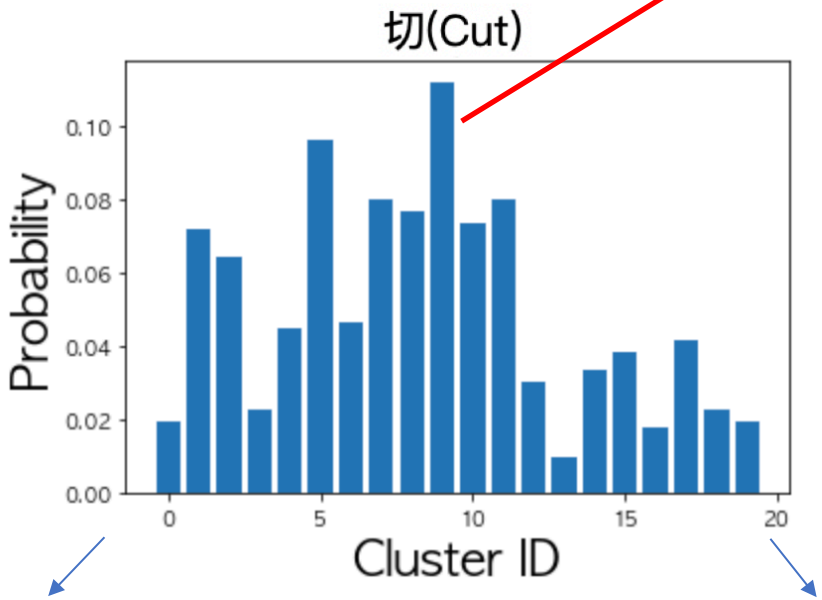


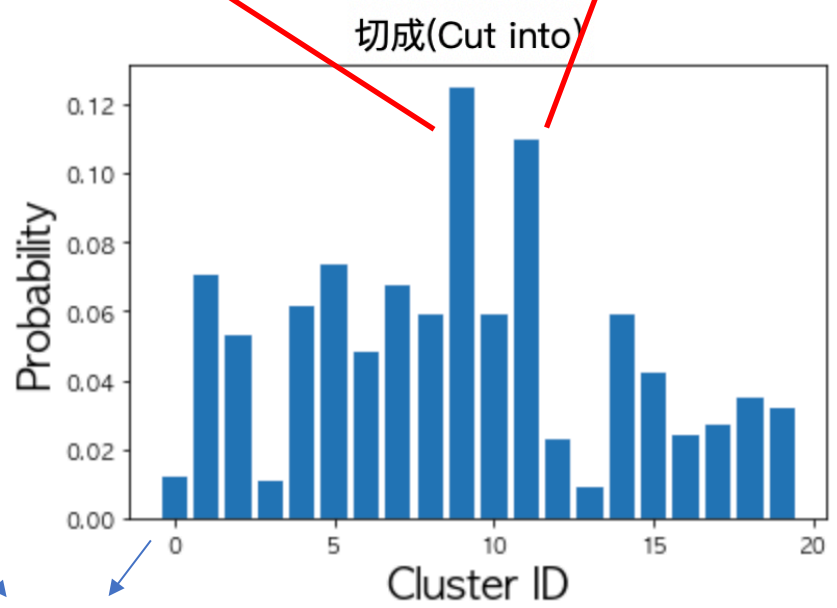
Proposed Method

- Probability distribution of associated images over clusters indicates the meaning of the verb.







Not similar

Similar

Dataset

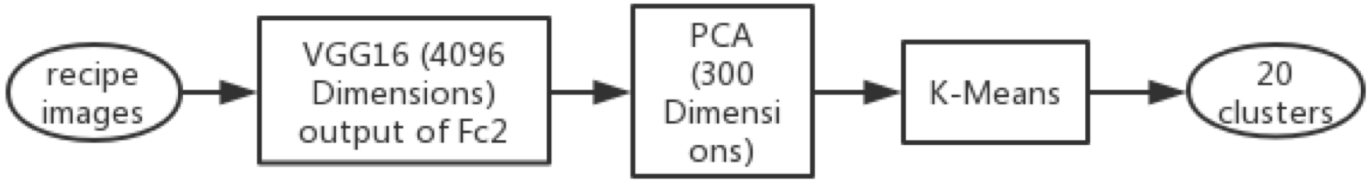
- #Recipes = 12,548
  - #Images = 48,164
  - Word segmentation and POS tagging by Jieba [1]
  - #Total Verbs = 269,993
  - #Distinct Verbs = 3,175
    - 10.7% appears 100 times or more.
    - 39.7% appears 2 to 9 times.
    - 21.9% appears only once.
    - Table 1 shows the top 20 most frequent words.

Table 1: Top 20 Most Frequent Verbs

rank	verb		frequency
1	put in	放入	16958
2	add in	加入	12187
3	pour in	倒入	7150
4	stir fry	翻炒	5413
5	prepare	准备	4792
6	boil	煮	4625
7	stir	搅拌	4613
8	set aside	备用	4590
9	moderate amount	适量	4341
10	wash	洗净	4291
11	add	加	3716
12	put	放	3096
13	out	出	2918
14	cut into	切成	2763
15	cut	切	2733
16	be	是	2550
17	clean	清洗	2202
18	mix well	拌匀	2171
19	ferment	发酵	2119
20	cover	盖	2089

[1] https://github.com/fxsjy/jieba

Image clustering method



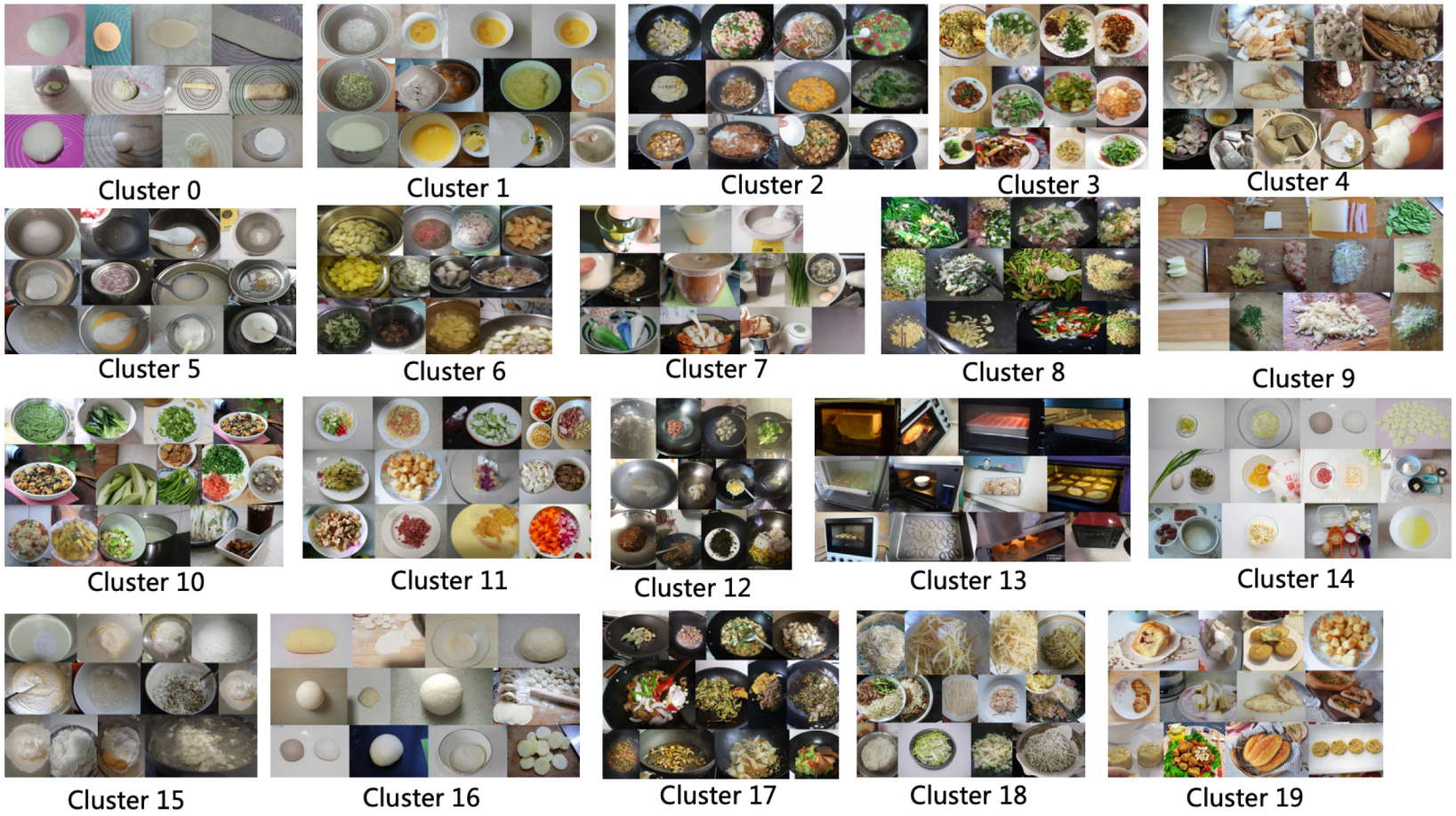
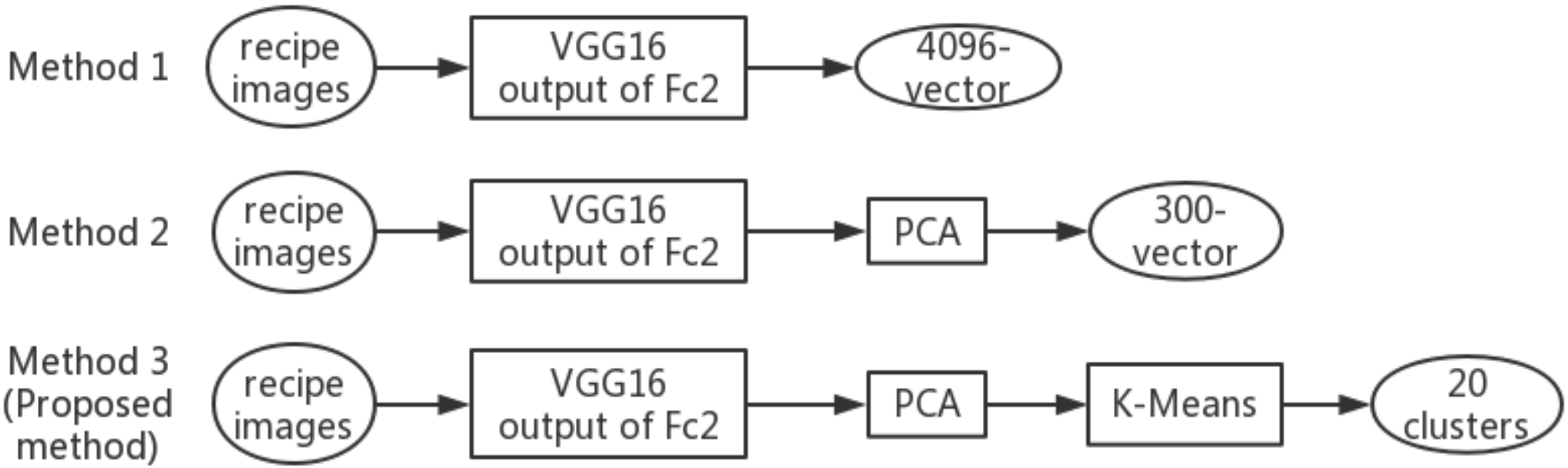


Table 1: The Number of Images in Each Cluster

cluster	#img	cluster	#img	cluster	#img	cluster	#img
00	775	05	3107	10	3468	15	2387
01	3244	06	2574	11	2530	16	1882
02	2561	07	3919	12	1436	17	1685
03	2696	08	2258	13	1123	18	1175
04	2664	09	3529	14	2544	19	2607

Comparison of methods

- Ground truth: Word2vec
  - To evaluate whether the proposed method achieves word embedding which represents word semantics as word2vec does.
- We compared **three ways** to vectorize verbs.
  - Method 1 and 2 use the image vectors calculated from VGG16 and PCA.
  - Method 3 uses the clusters of recipe images (proposed method).



Experiment

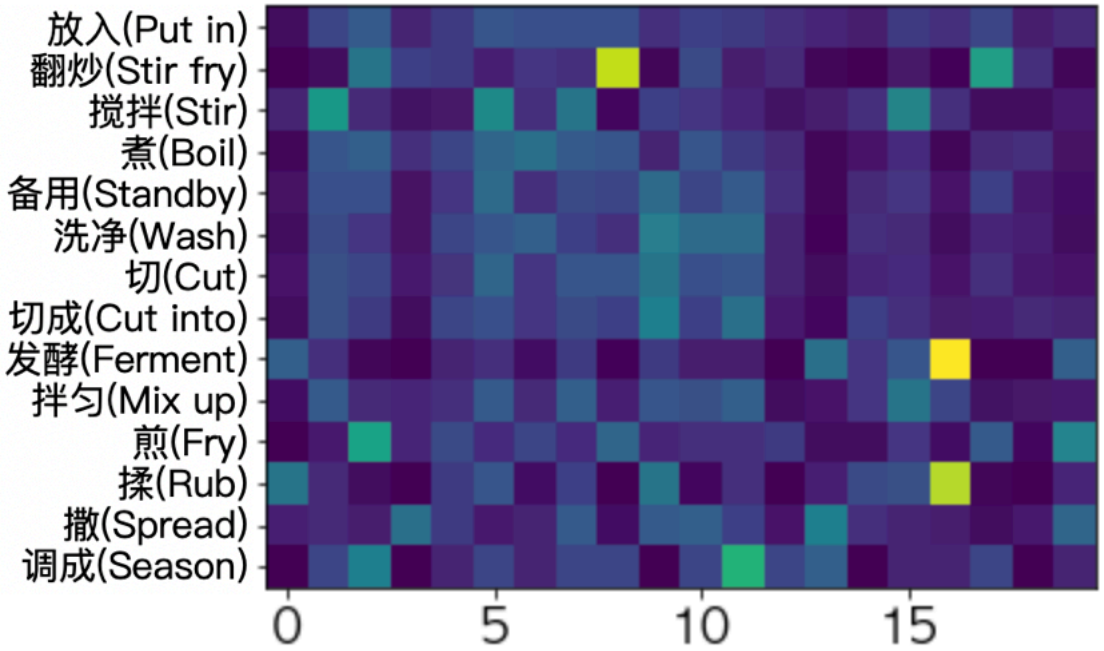
- Whether image styles contain semantics of associated image.
- Computing top-10 similar verbs by using our three vectorization methods.
- Calculating how many of the top-n results given by our image-based method are also included in the top-10 results given by the word2vec method for n = 1, . . . , 10.
- Dimensional compression from 4096 to 20 using image clustering did not cause performance deterioration.

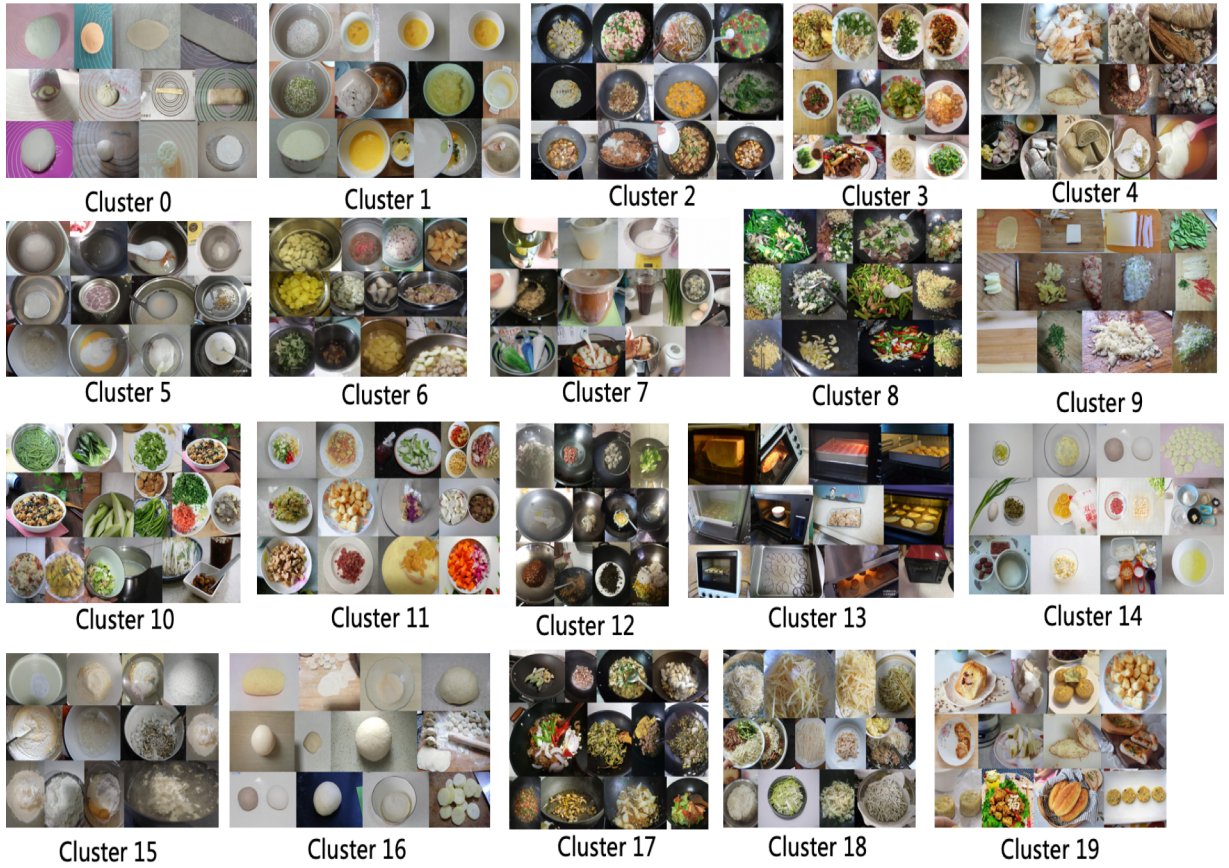
Table 5: Degree of Agreement with Text-Based Method

n	1	2	3	4	5	6	7	8	9	10
4096-vector	10	13	16	21	23	27	29	31	33	35
300-vector	10	13	17	21	24	27	29	31	35	39
20-vector	8	12	18	22	24	26	30	31	32	34

Experiment

- Producing a 20-dimensional vector for each verb by computing the ratio of each cluster within the set of all images associated with it.
- The heat map represents the ratio of each cluster for 14 example verbs.





Experiment

- The top-10 results of the word2vec method and the method using the 20-dimensional vectors for the 14 example verbs.
- Text-based (Word2vec) similarity and image-based (proposed method) similarity of verbs are complementary to each other.

boil 煮

	text		image	
1	boil	煮开	put in	放入
2	boil	煮沸	boil	煮沸
3	make soup	煲	boiling water	开水
4	cooked	煮熟	fish out	捞出
5	stew	焖	pour	倒入
6	boil	烧开	add in	加入
7	stew	炖煮	cook	煮熟
8	stew	炖	cooking wine	料酒
9	make soup	熬	stew	焖
10	turn	转	chop	切碎

Conclusion

- Word embedding of action verbs using pairs of text and images in recipes.
- Three methods of vectorizing verbs are compared with Word2vec.
  - Image-based word embedding achieves as good performance as Word2vec.
  - Dimensional compression from 4096 to 20 using image clustering did not cause performance deterioration.
  - Text-based similarity and image-based similarity of verbs are complementary to each other.

- Future work
  - To evaluate our proposed method by manually generated ground truth.
  - To further research on the relevance between textual and visual data.
    - Not only verbs, but also extend our method to other part-of-speech (e.g. adj.)
  - To deploy our method for recipe auto-translation or recipe multi-lingual retrieval.

Thank you