Categorization of Cooking Actions Based on Textual/Visual Similarity

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

Add less baking soda powder to the shrimp and a little salt and mix well, then wash with water, drain, then add the cooking wine and marinate for 10 minutes, then add starch and a little oil to grab.

source: https://www.haodou.com/recipe/1190778
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. Cut eggplant into sticks.
   - Wash the green pepper and cut.
   - Cut eggplant into sticks.

5. 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.

[1] https://github.com/fxsjy/jieba
Image clustering method

Table 1: The Number of Images in Each Cluster

<table>
<thead>
<tr>
<th>cluster</th>
<th>#img</th>
<th>cluster</th>
<th>#img</th>
<th>cluster</th>
<th>#img</th>
<th>cluster</th>
<th>#img</th>
<th>cluster</th>
<th>#img</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>775</td>
<td>05</td>
<td>3107</td>
<td>10</td>
<td>3468</td>
<td>15</td>
<td>2387</td>
<td></td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>3244</td>
<td>06</td>
<td>2574</td>
<td>11</td>
<td>2530</td>
<td>16</td>
<td>1882</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>2561</td>
<td>07</td>
<td>3919</td>
<td>12</td>
<td>1436</td>
<td>17</td>
<td>1685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>2696</td>
<td>08</td>
<td>2258</td>
<td>13</td>
<td>1123</td>
<td>18</td>
<td>1175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>2664</td>
<td>09</td>
<td>3529</td>
<td>14</td>
<td>2544</td>
<td>19</td>
<td>2607</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 \( n = 1, \ldots, 10 \).
- 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  |
• 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.
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.

<table>
<thead>
<tr>
<th>boil 煮</th>
<th>text</th>
<th>image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 boil</td>
<td>煮开</td>
<td>put in</td>
</tr>
<tr>
<td>2 boil</td>
<td>煮沸</td>
<td>boil</td>
</tr>
<tr>
<td>3 make soup</td>
<td>煲</td>
<td>boiling water</td>
</tr>
<tr>
<td>4 cooked</td>
<td>煮熟</td>
<td>fish out</td>
</tr>
<tr>
<td>5 stew</td>
<td>焖</td>
<td>pour</td>
</tr>
<tr>
<td>6 boil</td>
<td>烧开</td>
<td>add in</td>
</tr>
<tr>
<td>7 stew</td>
<td>炖煮</td>
<td>cook</td>
</tr>
<tr>
<td>8 stew</td>
<td>炖</td>
<td>cooking wine</td>
</tr>
<tr>
<td>9 make soup</td>
<td>熬</td>
<td>stew</td>
</tr>
<tr>
<td>10 turn</td>
<td>转</td>
<td>chop</td>
</tr>
</tbody>
</table>
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