A Case Study on Start-up of Dataset Construction: In Case of Recipe Named Entity Corpus

Yoko Yamakata, Keishi Tajima and Shinsuke Mori
Kyoto University, Japan
How We Start Dataset Construction?

- We have a new problem! Machine learning?
  → Need a new dataset! But it is an exhausting task…
  → Hiring annotators or crowdsourcing?
  → We need to give them a clear **annotation guideline**!
- Guideline creation typically goes like this…
  – Find similar datasets and learn their guidelines
  – Adapt them to our own tasks
  – Try annotation by ourselves to see if it is OK
  – Now let the annotators start to give annotation!

→ Even if you design the guideline very carefully, a lot of unexpected cases and ambiguity in rules will be found!
Three Main Problems

1. We need a system supporting the management of versions.
   • multiple versions of guidelines
   • also multiple versions of annotations under different guideline versions

2. How often should we update the guideline?
   • If very frequent many versions of the same annotation
   • If less frequent more annotation under old guidelines

3. When we have updated the guideline, which is better:
   • revising the old annotations under the new guideline, or
   • adding more data instead?
Our Task: Recipe Named Entity Corpus

Give tags of Recipe Named Entity (r-NE) to a cooking procedural text

10 types of r-NE

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Food</td>
</tr>
<tr>
<td>T</td>
<td>Tool</td>
</tr>
<tr>
<td>D</td>
<td>Duration</td>
</tr>
<tr>
<td>Q</td>
<td>Quantity</td>
</tr>
<tr>
<td>Ac</td>
<td>Action by chef</td>
</tr>
<tr>
<td>Ac2</td>
<td>Discontinuous Ac</td>
</tr>
<tr>
<td>Af</td>
<td>Action by food</td>
</tr>
<tr>
<td>At</td>
<td>Action by tool</td>
</tr>
<tr>
<td>Sf</td>
<td>Food state</td>
</tr>
<tr>
<td>St</td>
<td>Tool state</td>
</tr>
</tbody>
</table>

Example of annotation

Original text
Preheat oven to 200 C / Gas mark 6.

Annotation result
Preheat/Ac-B oven/T-B to/O 200/St-B C/St-I //O Gas/St-B mark/St-I 6/St-I ./O

24 min./recipe
Recipes were crawled at Allrecipes.co.uk

<table>
<thead>
<tr>
<th>dish type</th>
<th>#recipe</th>
<th>proportion</th>
<th>#corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread</td>
<td>953</td>
<td>3.1%</td>
<td>3</td>
</tr>
<tr>
<td>Pies and tarts</td>
<td>1251</td>
<td>4.0%</td>
<td>4</td>
</tr>
<tr>
<td>Soup</td>
<td>2046</td>
<td>6.6%</td>
<td>7</td>
</tr>
<tr>
<td>Salad</td>
<td>1755</td>
<td>5.7%</td>
<td>6</td>
</tr>
<tr>
<td>Main course</td>
<td>11523</td>
<td>37.2%</td>
<td>37</td>
</tr>
<tr>
<td>Dessert</td>
<td>3366</td>
<td>10.9%</td>
<td>11</td>
</tr>
<tr>
<td>Biscuits and</td>
<td>1655</td>
<td>5.3%</td>
<td>6</td>
</tr>
<tr>
<td>Pancakes</td>
<td>364</td>
<td>1.2%</td>
<td>1</td>
</tr>
<tr>
<td>Breakfast</td>
<td>1078</td>
<td>3.5%</td>
<td>3</td>
</tr>
<tr>
<td>Sandwiches</td>
<td>377</td>
<td>1.2%</td>
<td>1</td>
</tr>
<tr>
<td>Starters</td>
<td>2331</td>
<td>7.5%</td>
<td>8</td>
</tr>
<tr>
<td>Side dish</td>
<td>2166</td>
<td>7.0%</td>
<td>7</td>
</tr>
<tr>
<td>Sweets</td>
<td>416</td>
<td>1.3%</td>
<td>1</td>
</tr>
<tr>
<td>Preserves</td>
<td>423</td>
<td>1.4%</td>
<td>1</td>
</tr>
<tr>
<td>Drink</td>
<td>1231</td>
<td>4.0%</td>
<td>4</td>
</tr>
<tr>
<td>Cake</td>
<td>(4284)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>30935</td>
<td>100.0%</td>
<td>100</td>
</tr>
</tbody>
</table>

Recipes were crawled at Allrecipes.co.uk
Preparing the Guideline

1. We first defined tags and a guideline for Japanese recipe. http://www.ar.media.kyoto-u.ac.jp/how-to/recipe-NLP/

2. We translated them in order to adapt it to English recipe.

3. We hired a British doctor course student who had computational linguistics experience.

4. He soon sent us many questions!

- Even when we adopt an existing guideline, we have many unexpected cases!
- We had both
  - questions that do not require the revision of the guideline, and
  - many questions requiring discussions and guideline revisions!
P1: Prepositions and conjunctions are tagged O (i.e. outside an r-NE), except when they are part of a collocation.

P2: Adverbs and adverbial phrases are tagged O except when they are part of a phrasal verb.
- throw/Ac-B away/Ac-I
- mix/Ac-B in/O the/O bowl/T

P3: A sequence of words denoting a single action/food/tool in the cooking process is annotated as a single r-NE.
- frying/T-B pan/T-I
- bring/Ac-B to/Ac-I the/Ac-I boil/Ac-I

P4: Auxiliary and modal verbs are tagged O.

Examples of Questions

- “Pour/Ac-B into/O the/O digestive/F-B biscuit/F-I base F-I”
  Q. "digestive biscuit base" still part of the food?
  A. Yes

- “Butter/Ac-B 5/Q-B slices/F-B of/O bread/F-B”
  Q. Is “Butter” Ac (Action by chef) or F (Food)?
  A. Ac

- “Repeat/Ac-B with/O the/O other/Q-B dough/F-B balls/F-I ./O”
  Q. What’s the tag of “other”?
  A. Q (Quantity) revision of the guideline

- “fry/Ac-B diced/Ac-B bacon/F-B in/O a/O separate/St-B pan/T-B”
  Q. What’s the tag of “separate”?
  A. St (State of tool) revision of the guideline

- “continue/Ac-B cooking/Ac-I”
  Q. "continue cooking" a single NE?
  A. Yes revision of the guideline
### Examples of Revisions Adding New Tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Food</td>
<td>Eatables including intermediate products,</td>
</tr>
<tr>
<td>T</td>
<td>Tool</td>
<td>Knife, container, etc.</td>
</tr>
<tr>
<td>D</td>
<td>Duration</td>
<td>Duration of cooking</td>
</tr>
<tr>
<td>Q</td>
<td>Quantity</td>
<td>Quantity of food</td>
</tr>
<tr>
<td>Ac</td>
<td>Action by chef</td>
<td>Verbs representing of a chef’s actions</td>
</tr>
<tr>
<td>Ac2</td>
<td>Discontinuous Ac</td>
<td>words that consists single “Ac” with adjacent but not contiguous “Ac”.</td>
</tr>
<tr>
<td>Af</td>
<td>Action by food</td>
<td>Verbs representing food’s actions</td>
</tr>
<tr>
<td>At</td>
<td>Action by tool</td>
<td>Verbs representing tool’s actions. <strong>English only</strong></td>
</tr>
<tr>
<td>Sf</td>
<td>Food state</td>
<td>Food’s initial or intermediate states</td>
</tr>
<tr>
<td>St</td>
<td>Tool state</td>
<td>Tool’s initial or intermediate states</td>
</tr>
</tbody>
</table>

Addition of these tags needed deep discussions with experts in NLP and ML.
1) The requester describes the annotation guideline and sends it to annotators.

2) The annotators annotate data according to it and

3) return questions and exceptional cases that are not clearly specified in the guideline.

4) The requester discusses with supervisors if required,

5) revises the guideline from ver. $N$th to $(N + 1)$th, and

6) sends the revised guideline to the annotators.

7) The annotators update the previous annotation results to fit the current guideline.
Problem 1: Supporting Version Management

Guideline update brings different status of data which was annotated under different version of the guideline.
Problem 2: How often updating the guideline?

• We updated the guideline only twice.
  – We did not know which strategy (frequently or not) is better.
  – We were afraid of having repeated updates of the same annotation, which is inefficient.

• We cannot know if repeated updates could occur if we updated the guideline more often.

• Let us guess by looking at what types of revisions of annotations we needed.
Problem 2: How often updating the guideline?

Recipe: 100
Total word: 13,820
Tagged word: 7,107 (51.4)
Wrong tagged word: 2,584 (18.7%)

See next slide
Problem 3: Revising or Adding Annotations?

When we have updated the guideline, which is better?

- Revising old annotations under the old guideline (consistency of data)
- Adding more annotations by using the human resource for it (size of data)

Experiment:

Compare annotation accuracies by three types of training data

1. **First** annotation based on the initial guideline
2. **Final** annotation based on the final guideline
3. **50%-50% mixture** of these two

- Training data size (#recipes): 25, 50, 75, 100
- Test data: another 120 recipes
- Accuracies were evaluated with the named entity recognizer PWNER [Sasada et al. 2015]
Classification accuracy with different size of first and final annotation.

<table>
<thead>
<tr>
<th>#recipes</th>
<th>Final</th>
<th>Beginner</th>
<th>50%+50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>73%</td>
<td>68%</td>
<td>73%</td>
</tr>
<tr>
<td>50</td>
<td>75%</td>
<td>69%</td>
<td>75%</td>
</tr>
<tr>
<td>75</td>
<td>77%</td>
<td>71%</td>
<td>77%</td>
</tr>
<tr>
<td>100</td>
<td>78%</td>
<td>72%</td>
<td>78%</td>
</tr>
</tbody>
</table>

25 of final annotation > 100 of initial annotation → You should revise the guideline immediately!
Classification accuracy with different size of first and final annotation.

50 of final = 100 of 50%-50% mixture
If you have already 50 annotation results and have revised the guideline, you should update the old annotations rather than obtaining another 50 annotations.

F-measure

#recipes

65% 70% 75% 80%

68% 69% 71% 72%

73% 75% 77% 78%

73% 75% 77% 78%

80% 82%
Conclusion

• We need a system supporting management of versions of guidelines and versions of annotations under them.

• We should update the guideline frequently.
  – Repeated updates of the same annotations may not occur
  – Annotations under immature guidelines are quite unreliable

• When we have updated the guideline, we should revise the old annotations rather than adding more data.