Annotation Guidelines for Interactive Clustering of Cooking Recipe Instructions

Filippos Ventirozos, Mauricio Jacobo-Romero, Sarah Clinch and Riza Batista-Navarro
Department of Computer Science, University of Manchester
Manchester, United Kingdom
filippos.ventirozos@postgrad.manchester.ac.uk,
{*mauricio.jacoboromero,sarah.clinch,riza.batista*}*@manchester.ac.uk*

A. Introduction

This document describes a manual annotation task concerned with labelling sentences—and clusters thereof—drawn from cooking recipes. These sentences are constrained to those that contain instructions on how kitchen devices should be operated or configured. The labels assigned indicate a score: the extent to which pairs of items are semantically similar in terms of the types of events that they contain.

B. Kitchen Devices

In this study, we are focussing on two kitchen devices of interest: the oven and the fridge. However, other devices whose mechanical functionalities are similar to those of the oven and the fridge are also being considered, since we expect that they are involved in the same types of events. For the annotator's convenience, these other devices are listed in Table I under the "Subclasses" column. Meanwhile, terms which can be possibly confused with the oven and the fridge but are considered as irrelevant, are listed under the "Exceptions" column.

Table I
KITCHEN DEVICES OF INTEREST. INSTRUCTIONS INVOLVING ANY OF THE OVEN OR FRIDGE SUBCLASSES ARE RELEVANT; THOSE INVOLVING THE EXCEPTIONS ARE IRRELEVANT.

Oven		Fridge	
Subclasses	Exceptions	Subclasses	Exceptions
microwave	grill	refrigerator	cooling rack
toaster oven	dutch oven	freezer	
rotisserie (oven)		deep freezer	
furnace			

C. Event Types

Below we enumerate the types of kitchen device-related events which are of interest to our study (also depicted in Figure 1).

1 - Switch on (the device)
2 - Switch off (the device)
3 - Open the door (of the device)
4 - Close the door (of the device)
5 - Set the temperature (to any value, using any unit of measurement, e.g. Celsius or Fahrenheit)
6 - Set the timer (to any value, using any unit of measurement, e.g. minutes or hours)
7 - Set (device) mode (or functionality): activating any of various possible functionalities, e.g. "broiler", "grill", "convection" for an oven
8 - Set rack position, e.g. low, middle or high.
9 - Check appearance: assessing the food inside the device based on how it looks, e.g. "when golden brown"
10 - Check consistency: assessing the food inside the device based on its consistency, e.g. "until toothpick comes out clean"
11 - Rotate the food inside the device

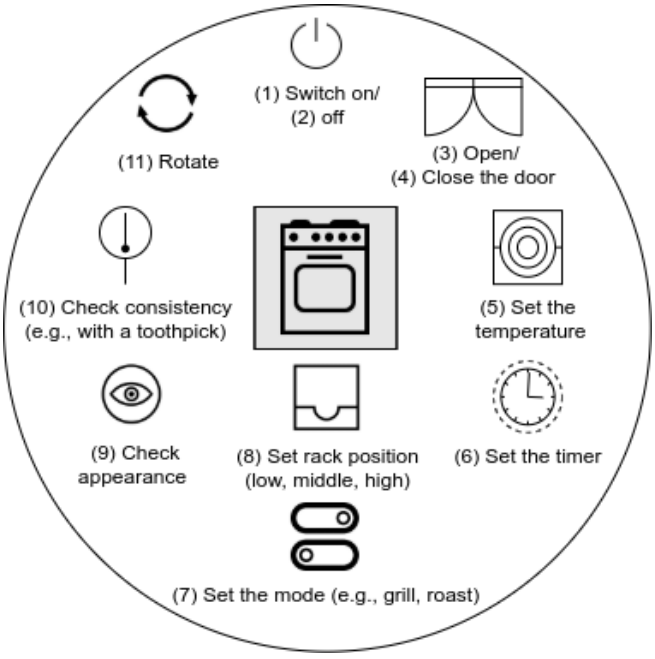

Figure 1. The event types covered in the annotation task. All 11 types apply to the oven; types 3, 4, 6, 9, 10 and 11 apply to the fridge.

We refer to the above 11 types as our typology of device operation events. In a sentence, more than one event type may apply.

D. Scoring

Drawing inspiration from the Semantic Textual Similarity (STS) task of SemEval 2017, we employ a scoring system whereby a score ranging from 0 to 5 is assigned to a sentence pair based on the following:

- **5**: if both sentences share the same device and the same event types
- **4**: if both sentences share the same device and more than half of the event types
- **3**: if both sentences share the same device and half of the event types
- **2**: if both sentences share the same device and less than half of the event types
- **1**: if both sentences do not share any event types but they share the same device
- **0**: if the sentences do not share the same device

We note that the following boundary cases might be encountered by the annotator:

- One sentence pertains to any of our kitchen devices of interest (the oven or fridge, or any of its subclasses): if they share the same event types, they can be assigned a score of **3**.
- Both sentences do not pertain to any kitchen device of interest: if the sentences describe the same types of events—even those which are not covered by our 11 event types, the annotator should use his/her discretion to assign a score.

E. Annotation Steps

The annotation procedure consists of three steps, depicted in Figures 2, 3 and 4 below.

As part of intra-cluster consistency checking (Figure 2), the annotator needs to judge whether all of the sentences contained in a given cluster are semantically similar. To this end, the annotator should pair up each sentence with every other sentence, and apply the scoring system described previously.

For inter-cluster consistency checking (Figure 3), the annotator needs to judge whether the two sentence clusters presented are semantically similar. Each cluster has already been judged as consistent in the previous step, hence the annotator needs to choose only one sentence from each cluster to form a sentence pair. The scoring system can then be applied to this pair.

Lastly, in outlier sentence checking (Figure 4), the annotator needs to pair up the given outlier sentence with any one sentence in the provided consistent cluster, and choose a score based on the scoring system.

F. Nuances

In this section, we provide some guidance on how the annotator should handle nuances in device operation events.

1) Variations in level of detail: Often, sentences are semantically similar in terms of the event types they describe, even if they vary in terms of level of detail.

- Only device-related events should be compared; other events should not form the basis for comparison. Hence, "take the food out of the oven" is the same as "with an oven mitt take the top part of the food out of the oven and sprinkle with spices" both refer to event type 3 (opening the device door). The same holds for "take it out of the oven" and "cool it on cooling racks when taken out of the oven".
- Conditioned events are also used for comparison. For instance, "if you want store it in the fridge" is the same as "store it in the fridge".

2) Differences in purpose: Sometimes, events are described in sentences using the same linguistic constructions, although the instructions are for completely different purposes. In such cases, the annotator should consider the sentences as sharing half of the event types (score of **3**).

- The instruction "microwave for 30 s" is not the same as "microwave at 30 s intervals". They share event type 6, but the latter instruction is specifically for operating the device at intervals.
- The instruction "put it in the fridge for 2 hours" is not the same as "you can store it in the fridge for a month", as the latter is specifically for the purpose of long-term storage.

3) Implied events: In some cases, a sentence might describe an event type that is only implied. The annotator should consider the implied event type in their comparison.

- The instruction "bake it until done" refers to event type 9 since it is implied that appearance will be checked to assess that the food is done.
- In "once the food is out of the oven sprinkle it with spices", event type 3 applies since it is implied that the device door needs to be opened in order to take the food out of the oven.
- In "bake the food for 20 mins", it is implied that the device door needs to be opened to put the food in, then it needs to be closed, and that the timer needs to be set. Hence event types 3, 4 and 6 apply to this case.

If the annotator finds difficulty in understanding some technical cooking terms, they can refer to a comprehensive glossary of common culinary terms[1].

[1]https://ueat.utoronto.ca/the-student-kitchen/kitchen-basics-techniques/glossary-common-culinary-terms/

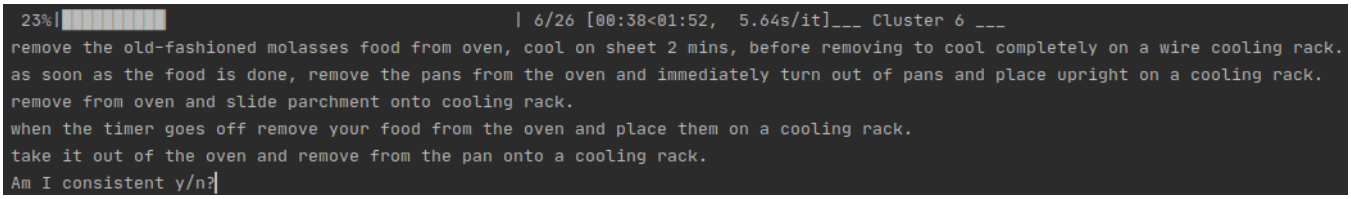

Figure 2. Intra-cluster consistency checking. The annotator should choose “y” if all of the sentences in a cluster are semantically similar.

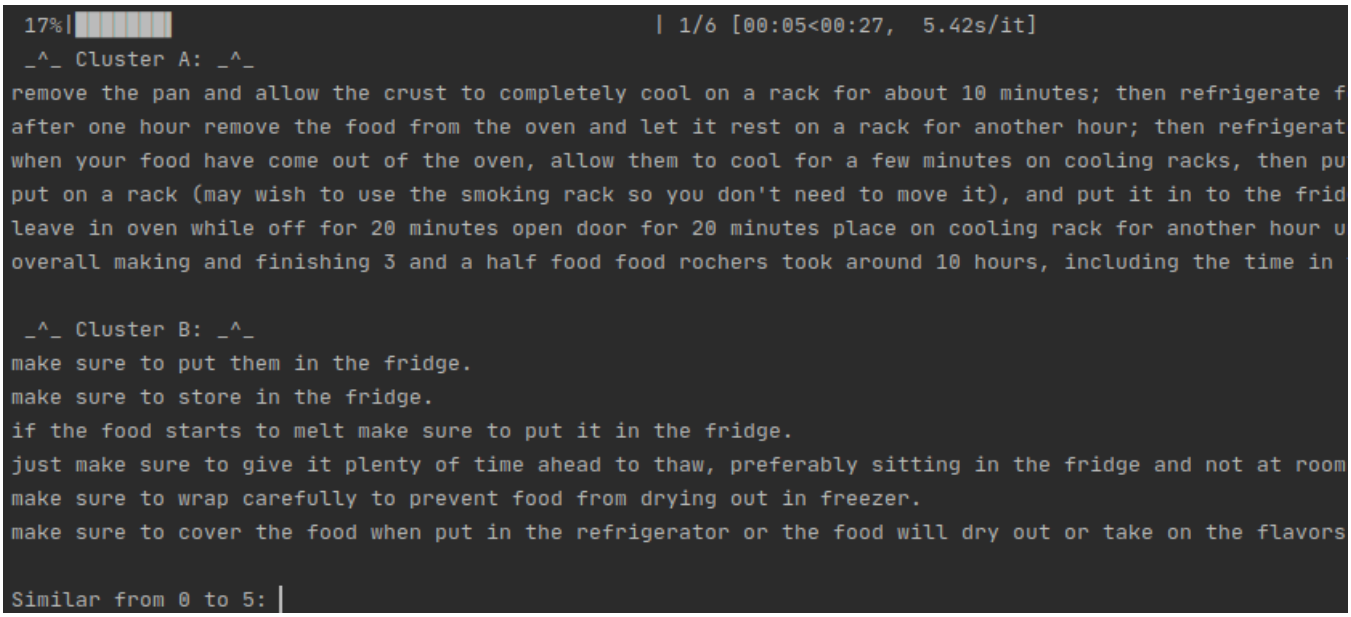

Figure 3. Inter-cluster consistency checking. The annotator should assign a score ranging from 0 to 5, based on the semantic similarity between the two given consistent clusters.

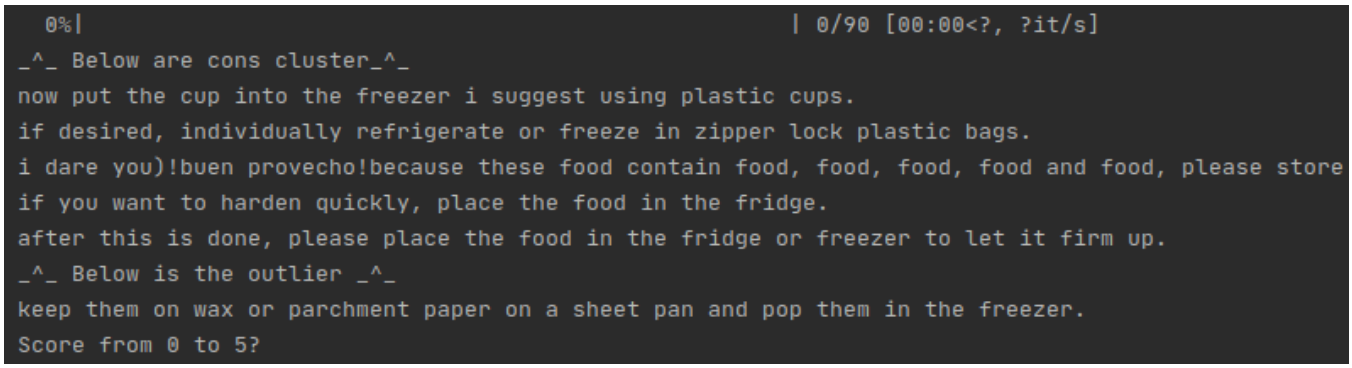

Figure 4. Outlier sentence checking. The annotator should label the provided outlier sentence with a score ranging from 0 to 5, based on its semantic similarity to the given cluster.