Food Concept Mapping Annotation Guidelines

1. Introduction

The purpose of this document is to provide a standard to create labels for food terminology mapping. The ontology for which we will map towards is Food Ontology (FoodOn). The annotated corpus will serve as the golden standard for food entity extraction and mapping towards FoodOn.

The corpus we will be annotating is meal data obtained from Dr. Lena Mamykina. Meal data are manually curated by patients or consumers of the application. Misspellings, abbreviations, and ambiguous expressions are common within such data. Here we will standardize the way of mapping non-trivial terminology and provide examples of how they're done.

2. What are food entities?

Here we will loosely define food entities as words, terms, or phrases that are commonly understood as a food or agricultural product. This includes terms represented by brand names, e.g., Big Mac, acronyms, e.g., PBJ, and colloquial language, e.g., greens. Not all identified food entities will be included within recognized food ontology (i.e., there will not be 1-to-1 mapping between identified entities and FoodOn IDs.)

3. How concepts are mapped to FoodOn codes

We will select the label that represents the food entity with greatest granularity. In other words. For example, greens will be tagged with ID for salad, while greens with chicken added will be tagged with ID for chicken salad.

We will not provide additional tags for misspellings. Misspelled food entities will be tagged with their respective FoodOn IDs and treated as entities with accurate spelling.

Acronyms and abbreviations that are abbreviated for foods are expanded and tagged with ingredients of the food entity. For example, BLT will be tagged with bacon, lettuce, tomato, and sandwich. Such acronyms that do not exist within FoodOn will receive an additional tag FOOD_ABAC.

Branded items that do not exist within FoodOn will be given a unique tag FOOD_BRAND.

For ambiguous food entities that require context for identification, an additional tag of FOOD_AM is provided. For entities that cannot be determined given context, we will remove them from the cohort.

All tokens that represent food will be annotated even if the term is repeated or general.

Sample annotations:

1. Food entity represented by single word
    a. Pretzel (FOOD_N)
2. Food entity represented by multiple words
    a. Mixed vegetables (FOOD_N), and salad
3. Food entity represented by non-sequential terms
    a. Honey (Modifier) bunches of oats (FOOD_AM)
    Here we will consider the entity as honey oats. The two terms will be "linked" as a single entity. Modifier and FOOD_AM will be connected by the relationship "Contain".
4. Food entity that are misspelled
    a. Tomatoes salad chiken cutlets (FOOD_N)
5. Food entity represented by acronym
    a. BBQ (FOOD_ABAC) ribs
6. Food entity represented by abbreviation
    a. Sub (FOOD_ABAC) with tomatoes
7. Food entity represented by brand names
    a. Big Mac (FOOD_BRAND), fries and coke (FOOD_BRAND).
8. Food entity that is ambiguous or require context for identification
    a. Beans (FOOD_AM). Red beans (FOOD_N), steam spinach, baked chicken.
    Here we can interpret the first mention of "Beans" to *bean plant* under *pod or seed vegetable plant*
9. Food entity that is removed
    a. Rice and beans.
    We do not know whether it is green beans or black beans for this entity.
10. Food entity that is removed
    a. Rice and beans.
    We do not know whether it is green beans or black beans for this entity.
11. Example annotation of an entire meal entry
    a. Vegetable (FOOD_N), spinach (FOOD_N) and carrots (FOOD_N).
    All terms that are identified as food should be annotated.

4. Tool for annotation

We selected Brat as the annotation tool for this study (https://brat.nlplab.org/). FOOD_N, FOOD_ABAC, FOOD_BRAND, FOOD_AM, Modifier are declared as entity types. Contain is declared as a relationship. FoodOn IDs are recorded within the note section of each food concept. There are some limitations associated with this structure. For example, one concept can only be tagged with one entity type.